REMARKS

Claims 1 to 15 and 17 to 27 are pending.

Election

The Examiner has withdrawn claim 17 from consideration, on the basis that it is drawn to a non-elected species. The Applicants respectfully submit that claim 17 is drawn to an elected species. Specifically, in the previous response the Applicants provisionally elected Species II-A, which as defined by the Examiner in the previous Office Action includes: "[m]ethods as set forth in Group I, II, and III, wherein said set of characteristics comprises non-genetic factors" (Emphasis added). Claim 17 recites "said set of characteristics comprises both genetic and non-genetic factors" (Emphasis added). Thus, Claim 17 is drawn to Species II-A and re-consideration and examination of Claim 17 is respectfully requested.

Information Disclosure Statement

The Examiner stated that the Schafer et al. reference included in the submitted Information Disclosure Statement has not been considered as a legible copy was not available. A legible copy of the relevant portion of Schafer et al. including the cover pages and Chapter 9, pp. 333-377, is enclosed for reconsideration by the Examiner.

Objection

Claims 1, 4, 22 and 23 have been amended to delete the numbers with trailing periods, in response to the Examiner's objections.

Statutory Subject Matter

The Examiner rejected claims 1 to 15 and 18 to 27 under 35 U.S.C. §101 as being directed to non-statutory subject matter. In particular, the Examiner is of the view that claim 1 does not produce an actual, concrete result in a tangible form useful to one skilled in the art. The Applicants respectfully traverse this rejection.

As outlined in the "Interim Guidelines for Examination of Patent Applications for Patent Subject Matter Eligibility" – OG Date, 22 November 2005 ("the Guidelines"):

- "If the claim is directed to a practical application ... producing a result tied to the physical world that does not preempt the judicial exception, then the claim meets the statutory requirement of 35 U.S.C. §101." (Emphasis added)
- "The tangible requirement does not necessarily mean that a claim must either be tied to a particular machine or apparatus or must operate to change articles or materials to a different state or thing. ...the process claim must set forth a practical application ... to produce a real-world result.
- "the opposite meaning of 'tangible' is 'abstract'."
- "the opposite of 'concrete' is unrepeatable or unpredictable."
- "If the record as a whole suggests that it is more likely than not the claimed invention would be considered a practical application of an abstract ideal, natural phenomenon, or law of nature, the Examiner should not reject the claim." (Emphasis added)

Each of Claim 1 and claim 21 recites, among others, determining a plurality of weights associated with collected sets of data, each associated with a member of a population, and optimizing the parameters of a candidate statistical model, taking into account of the weights. The weights and optimized parameters can be repeatedly and predictably produced, to create a statistical model for predicting disease risk for a member

of the population. The results s are thus are not merely "abstract" numbers but are useful, concrete and tangible results that have practical application in the "real world". For example, the weights indicate the statistical significance of data sets, which are useful, for optimizing risk prediction models which can be used for calculating disease risks to members of a specific population (claim 1), as described in the present application.

In addition, Claim 1 has been amended to clarify that the parameters of the chosen risk model are optimized so that a risk calculated using the risk model and a set of data of the first type associated with a particular member of the population is indicative of a disease risk to the particular member, which is a further practical application of the risk model in the real world. Support for the amendment can be found in the description at, e.g., paragraphs [0094] and [00124], and in FIGS. 2 and 7.

Claim 21 has been similarly amended to recite that an optimized risk model is stored for use in calculating disease risks. Support for storing statistical models can be found in the description, such as at paragraph [0039], and in FIG. 1.

Claims 2 to 15, 17 to 20, and 22 to 26 depend from one or other of claims 1 and 21 directly or indirectly.

It is submitted that the record as a whole, including the description, clearly sets forth how the claimed invention involves practical applications of the determined weights or risk model, and the claims on file do not preempt any of the judicial exceptions. Therefore, it is believed that the current claims 1 to 15 and 17 to 26 are directed to statutory subject matter and withdrawal of rejections of these claims on the basis of non-statutory subject matter is respectfully requested.

The Examiner did not articulate why claims 20 and 27 are considered to be directed to non-statutory subject matter. Claim 20 is directed to a computer system not a "method" as asserted by the Examiner. Examiner did not provide reasons for rejecting independent Claim 27 in sufficient detail so that the Applicants can properly respond. As the "Examiner

bears the initial burden ... of presenting a *prima facie* case of unpatentability" (see the Guidelines, IV - D.), the Examiner is therefore requested to either withdraw the rejections to claims 20 and 27 or to provide a sufficient basis for rejecting claim 20 or 27 so that the Applicants can better address the rejection.

§112 rejection

The Examiner further rejected Claims 1 to 15 and 22 to 26 under 35 U.S.C. §112, second paragraph.

In particular, the Examiner stated that it is unclear what steps are required to determine the statistical model in claim 1. Claim 1 has been amended to clarify that the candidate statistical model with the parameters optimized as claimed is chosen as the statistical risk model.

The Examiner further notes that the limitation "said data having like data of said second type" lacks a proper antecedent basis. In response, the Applicants note that the complete limitation is "sets of said data having like data of said second type", which is believed to be properly introduced and has proper antecedent basis. The Examiner also stated that the meaning of the word "like" is unclear. It is submitted that it would be clear to a person skilled in the art what "like data" means in the context of claim 1. It would be clear when the claim is read as a whole that "like data" are data that are similar or alike, possess similar characteristics, or have identical or equivalent values. It is thus believed that no further clarification is necessary. Withdrawal of this objection is requested.

Claim 3 has been amended to delete the word "corresponding", thus addressing the Examiner's objection on the basis of lack antecedent.

The Examiner further noted that the sentence "a reference group which contains sets of data having data of said second type like data of said second type obtained from said member of said population" in claim 4 is unclear as written as it is not clear in what way the

"reference group" is further limited. In response, the Applicants note that this sentence has a reasonable interpretation: as far as data of the second type is concerned, the data contained in the reference group and the data obtained from the member are alike, which is a further limitation on the reference group. Thus, it is believed that no further clarification is needed. Withdrawal of this objection is requested.

Claims 7, 13 and 27 have been amended to provide proper antecedents for all recited elements. It is believed that these amendments address the Examiner's objections to claims 7 and 13 under §112, second paragraph. The Applicants note that a common and ordinary meaning of the word "representativeness" is the quality or state of being representative. Read in the context of the claims as amended, a person skilled in the art would understand that this word refers to the extent to which the member is representative in the population.

As the objections to claim 1 under §112 have been addressed as discussed above, withdrawal of the objections to claims 2, 5, 6, 8 to12, 14, 15, and 23 to 26 as they depend either directly or indirectly on claim 1 is requested.

Prior Art Rejections

The Examiner rejected claims 1, 3, 8 to 11, 13 and 19 to 21 under 35 U.S.C. §102(b) as being anticipated by Schoonjans. The Applicants respectfully traverse the rejection.

Specifically, claim 1 recites, among others, collecting sets of data each set associated with one member of a population, selecting a candidate statistical model dependent on a plurality of parameters, determining weights each associated with one collected set of data, and optimizing the parameters by fitting the model to the collected sets of data, taking into account of the weights. The collected sets of data are not any data. They must include input data to be used in the fitting. Each weight indicates a statistical significance of its associated set of data.

In stark contrast, Schoonjans discloses a hazard model H(t) dependent on a collection of predictor variables (X) and coefficients (b) associated with the variables. The coefficients "b" are estimated by Cox regression. However, Schoonjans does not disclose or suggest determining weights associated with collected data sets used in the estimation of "b" by Cox regression. It apparently could not disclose optimizing the coefficients "b" by taking into account of the weights, as no such weights are determined. The Examiner takes the position that "exp(b)" are the "weights" as recited in claim 1. This position is incorrect. The coefficients "b" are associated with the predictor variables (X), which are covariates or risk factors (see page 1 of Schnoonjans). As any person skilled in the art would understand, these variables are not individually associated with individual members of a population. They are not the sets of data each associated with one member of the population, as defined in claim 1. Further, claim 1 calls for both parameters and weights, which are distinct and separate quantities. The coefficients "b" may be considered either as parameters or weights, but they cannot be both the parameters and the weights as defined in claim 1.

As Schoonjans does not disclose all of the limitations of claim 1, it is submitted that claim 1, and the claims dependent therefrom directly or indirectly, are not anticipated by Schoonjans.

The Examiner further rejected claims 1, 2, 4, and 10 under 35 U.S.C. §102(b) as being anticipated by Lloyd et al. The Applicants respectfully traverse the rejection.

Specifically, the Examiner asserts that Lloyd et al. teaches:

> Model includes fixed parameters to be estimated that is a multiplier of time dependent covariates associated with statistical significant [p.149, lines 23-27], which is a teaching for a "candidate statistical model' and "weights associated with statistical significance" as in instant claim 1(c).

This statement is incorrect. As discussed above, any person skilled in the art would understand neither the fixed parameters nor the covariates of a model are collected sets of data each associated with one member of the population as defined in claim 1. A review of Lloyd et al. indicates that they do not disclose or suggest determining weights each associated with a collected set of data where the set of data is associated with one member of a population. Thus, it is respectfully submitted that claim 1, and claims 2, 4 and 10 dependent therefrom directly or indirectly, are not anticipated by Lloyd et al.

Withdrawal of the rejections under 35 U.S.C. §102(b) is therefore respectfully requested.

The Examiner also rejected claims 1 to 3 under 35 U.S.C. §103(a) as obvious having regard to Kirchberg et al. in view of Montomoli et al. The Applicants respectfully traverse this rejection.

First of all, Kirchberg et al. is related to genetic model of optimization for Hausdorff distance-based face localization, which is in an art non-analogous to the art of disease risk predication. It would not have been obvious for a person skilled in the art of disease risk predication to look for references related to the art of face localization or image analysis.

Further, the Examiner relies on Kirchberg et al. for disclosing the limitations of former claim 1, including: "collecting ...sets of data, each...associated with one member of said population, and comprising...an indicator of disease status" [limitation 1(a)]; and "selecting a candidate statistical model for calculating said disease risk..." [limitation 1(b)]. Careful review of Kirchberg et al. reveals that Kirchberg et al. do not disclose or suggest any of these limitations as recited in claim 1.

Specifically, the Examiner stated that Kirchberg et al. teach "[d]evelopment of a 'face model' consisting of feature points [p.104, paragraph 4], as in instant claim 1(b)". This statement is incorrect. The "face model" disclosed in Kirchberg et al. is for locating and representing possible faces in an image (see p. 104, paragraphs 1 to 6 of Kirchberg et al.).

Kirchberg et al. do not disclose or suggest using the "face model" for calculating disease risks. Nor does Montomoli et al.

Further, the Examiner stated that Kirchberg teach "a metric..., which correlates to a[n] 'indicator' as in instant claim 1(a)". This is also incorrect. Kirchberg et al. do not disclose or suggest the collection of data sets each comprising "an indicator of disease status", as claimed in Claim 1. As recognized by the Examiner, the "metric" disclosed in Kirchberg is "for determining distance between two data points" in an image, which has nothing to do with the disease status of individual members of a population.

As the Examiner has failed to show the cited references, either alone or in combination, disclose all of the limitations of claim 1, or any of claims 2 and 3 dependent therefrom, the Examiner has failed to establish a *prima facie* case of obviousness.

Withdrawal of the rejection under §103(a) is therefore respectfully requested.

No new matter has been added by this amendment.

In view of the foregoing, favourable consideration of the application is respectfully requested.

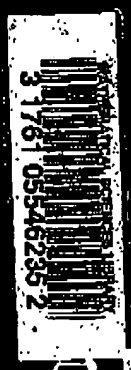
Respectfully submitted,

DINSMORE & SHOHL, LLP

By _____
Timothy W. Hagan
Registration No. 29,001

One Dayton Centre
One South Main Street, Suite 1300
Dayton, Ohio 45402
Tel. (937) 449-6400
Fax (937) 449-6405

August 28, 2006
Enclosures
MZ/JJP/jkk
92706-41



Monographs on Statistics and Applied Probability 72

Analysis of Incomplete Multivariate Data

J. L. Schafer

CHAPMAN & HALL

Analysis of Incomplete Multivariate Data

J.L. SCHAFER

Department of Statistics
The Pennsylvania State University
USA

CHAPMAN & HALL
London · Weinheim · New York · Tokyo · Melbourne · Madras

Published by Chapman & Hall, 2-6 Boundary Row, London SE1 8HN, UK

Chapman & Hall, 2-6 Boundary Row, London SE1 8HN, UK

Chapman & Hall GmbH, Pappelallee 3, 69469 Weinheim, Germany

Chapman & Hall USA, 115 Fifth Avenue, New York, NY 10003, USA

Chapman & Hall Japan, ITP-Japan, Kyowa Building, 3F, 2-2-1 Hirakawacho, Chiyoda-ku, Tokyo 102, Japan

Chapman & Hall Australia, 102 Dodds Street, South Melbourne, Victoria 3205, Australia

Chapman & Hall India, R. Seshadri, 32 Second Main Road, CIT East, Madras 600 035, India

First edition 1997

Printed in Great Britain by St Edmundsbury Press, Bury St Edmunds, Suffolk

ISBN 0 412 04061 1

A catalogue record for this book is available from the British Library

Printed on permanent acid-free text paper, manufactured in accordance with ANSI/NISO Z39.48-1992 and ANSI/NISO Z39.48-1984 (Permanence of Paper).

[redacted]

CHAPTER 9

Methods for mixed data

9.1 Introduction

Chapters 5–8 pertained to datasets in which the variables were either all continuous or all categorical. In practice, however, statistical analyses involving variables of both types are extremely common: analysis of variance, analysis of covariance, logistic regression with continuous predictors, and so on. Sample surveys often contain variables of both types. This chapter develops general tools for incomplete multivariate data matrices containing both continuous and categorical variables. Such a dataset is shown in Figure 9.1, with missing values denoted by question marks.

The statistical literature on multivariate methods tends to emphasize models for variables that are all of the same type; relatively

	categorical				continuous			
units	1	W_1	W_2	...	W_p	Z_1	Z_2	... Z_q
	2				?	?		
	3				?			?
	.							
	.		?				?	
	.				?			?
	.		?					
	n	?				?		?
			?			?		

Figure 9.1. *Mixed dataset with missing values.*

little attention has been paid to models for mixed data. One notable exception is the model that underlies classical discriminant analysis, which contains a single categorical response and one or more continuous predictors. We begin with a version of this model called the general location model (Section 9.2) and discuss methods for keeping the number of parameters manageable (Section 9.3). Algorithms for incomplete mixed data are presented in Section 9.4, and Section 9.5 concludes with several data examples.

9.2 The general location model

9.2.1 Definition

As in Figure 9.1, let $W_1, W_2, \ldots, W_p$ denote a set of categorical variables and $Z_1, Z_2, \ldots, Z_q$ a set of continuous ones. If these variables are recorded for a sample of n units, the result is an $n \times (p+q)$ data matrix $Y = (W, Z)$, where W and Z represent the categorical and continuous parts, respectively.

The categorical data W may be summarized by a contingency table. Let us suppose that W_j takes possible values $1, 2, \ldots, d_j$, so that each unit can be classified into a cell of a p-dimensional table with total number of cells equal to $D = \prod_{j=1}^{p} d_j$. A generic response pattern for the categorical variables will be denoted by $w = (w_1, w_2, \ldots, w_p)$, and the frequencies in the complete-data contingency table will be

$$x = \{x_w : w \in \mathcal{W}\}, \tag{9.1}$$

where x_w is the number of units for which $(W_1, W_2, \ldots, W_p) = w$, and $\mathcal{W}$ is the set of all possible w. We may also arrange the cells of the contingency table in a linear order indexed by $d = 1, 2, \ldots, D$, for example, the anti-lexicographical storage order in which w_1 varies the fastest, w_2 varies the next fastest, and so on (Appendix B). Then we can replace the vector subscript in x_w by a single subscript d,

$$x = \{x_d : d = 1, 2, \ldots, D\}. \tag{9.2}$$

Depending on the context, we will regard x either as a multidimensional array (9.1) or a vector (9.2).

Finally, it will be helpful to introduce one additional characterization of W. Let U be an $n \times D$ matrix with rows u_i^T, $i = 1, 2, \ldots, n$, where u_i is a D-vector containing a 1 in position d if unit i falls into [illegible] and 0s in all other positions. Hence each row of U contains

a single 1, and U^TU is

$$U^TU = \operatorname{diag}(x) = \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_D \end{bmatrix}. \tag{9.3}$$

Because the sample units are assumed to be independent and identically distributed, all relevant statistical information in W is contained in x, U or U^TU. The continuous data are characterized simply by Z.

The general location model, so named by Olkin and Tate (1961), is most easily defined in terms of the marginal distribution of W and the conditional distribution of Z given W. The former is described by a multinomial distribution on the cell counts x,

$$x \mid \pi \sim M(n, \pi), \tag{9.4}$$

where $\pi = \{\pi_w : w \in \mathcal{W}\} = \{\pi_d : d = 1, 2, \ldots, D\}$ is an array of cell probabilities corresponding to x. Given W, the rows $z_1^T, z_2^T, \ldots, z_n^T$ of Z are then modeled as conditionally multivariate normal. Let E_d denote a D-vector containing a 1 in position d and 0s elsewhere. We assume

$$z_i \mid u_i = E_d, \mu_d, \Sigma \sim N(\mu_d, \Sigma) \tag{9.5}$$

independently for $i = 1, 2, \ldots, n$, where μ_d is a q-vector of means corresponding to cell d, and Σ is a $q \times q$ covariance matrix. The means of $Z_1, Z_2, \ldots, Z_q$ are allowed to vary freely from cell to cell, but a common covariance structure Σ is assumed for all cells. When $D = 2$, this reduces to the model that underlies classical discriminant analysis (e.g. Anderson, 1984).

The parameters of the general location model will be written

$$\theta = (\pi, \mu, \Sigma),$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_D)^T$ is a $D \times q$ matrix of means. For the moment, we will impose no prior restrictions on θ other than the necessary positive definiteness for Σ and $\sum_{w \in \mathcal{W}} \pi_w = 1$. The number of free parameters in the unrestricted model is thus

$$(D-1) + Dq + q(q+1)/2.$$

Notice that the model for Z given W may also be regarded as a classical multivariate regression,

$$Z = U\mu + \epsilon \qquad (9.6)$$

where ϵ is an $n \times q$ matrix of errors whose rows are independently distributed as $N(0, \Sigma)$. The columns of U contain dummy variables for each of the cells $d = 1, 2, \ldots, D$. Because U has the same rank as $U^T U = \mathrm{diag}(x)$, this will be a full-rank regression provided that there are no random zeroes in x. Structural zeroes may be handled simply by omitting them from the columns of U. A model of the form (9.6) is sometimes called a *standard multivariate regression*; in this model the same matrix of regressors U is used to predict each column of the response Z.

9.2.2 Complete-data likelihood

Combining (9.4) with (9.5), we can write the complete-data likelihood as the product of multinomial and normal likelihoods,

$$L(\theta \mid Y) \propto L(\pi \mid W)\, L(\mu, \Sigma \mid W, Z). \tag{9.7}$$

The likelihood factors are $L(\pi \mid W) \propto \prod_{d=1}^{D} \pi_d^{x_d}$ and

$$L(\mu, \Sigma \mid W, Z) \propto |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\tfrac{1}{2} \sum_{d=1}^{D} \sum_{i \in B_d} (z_i - \mu_d)^T \Sigma^{-1} (z_i - \mu_d) \right\},$$

where $B_d = \{i : u_i = E_d\}$ is the set of all units belonging to cell d. After some algebraic manipulation, the second factor may be written as

$$L(\mu, \Sigma \mid W, Z) \propto |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\tfrac{1}{2} \mathrm{tr}\, \Sigma^{-1} Z^T Z + \mathrm{tr}\, \Sigma^{-1} \mu^T U^T Z - \tfrac{1}{2} \mathrm{tr}\, \Sigma^{-1} \mu^T U^T U \mu \right\}, \tag{9.8}$$

revealing that the complete-data loglikelihood is linear in the elements of the sufficient statistics

$$T_1 = Z^T Z, \quad T_2 = U^T Z, \quad \text{and} \quad T_3 = U^T U = \mathrm{diag}(x). \tag{9.9}$$

Maximum-likelihood estimates

Because the parameters associated with the two factors in (9.7) are distinct, complete-data ML estimates may be found by maximizing each factor separately. The result for π is the usual ML estimate for an unrestricted multinomial model,

$$\hat{\pi}_d = \frac{x_d}{n}, \quad d = 1, 2, \ldots, D.$$

The estimate for μ follows from the least-squares regression of Z on U,

$$\hat{\mu} = (U^T U)^{-1} U^T Z = T_3^{-1} T_2, \tag{9.10}$$

and the estimate for Σ is

$$\hat{\Sigma} = \frac{1}{n} \hat{\epsilon}^T \hat{\epsilon} = \frac{1}{n} \left(T_1 - T_2^T T_3^{-1} T_2 \right), \tag{9.11}$$

where $\hat{\epsilon} = Z - U\hat{\mu}$ is the matrix of estimated residuals. Notice that (9.11) differs from the classical unbiased estimate in that it uses a denominator of n rather than $n - D$.

These estimates can be further understood by noting that

$$(U^T U)^{-1} = \begin{bmatrix} x_1^{-1} & 0 & \cdots & 0 \\ 0 & x_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_D^{-1} \end{bmatrix}$$

and that $U^T Z$ is a $D \times q$ matrix with $\sum_{i \in B_d} z_i^T$ in the dth row. The rows of $\hat{\mu}$ are thus

$$\hat{\mu}_d^T = x_d^{-1} \sum_{i \in B_d} z_i^T, \quad d = 1, 2, \ldots, D,$$

the within-cell averages of the rows of Z. The rows of the residual matrix $\hat{\epsilon}$ are the deviations of the rows of Z from their cell-specific means, so the estimated covariance matrix can be written as

$$\hat{\Sigma} = \frac{1}{n} \sum_{d=1}^{D} \sum_{i \in B_d} (z_i - \hat{\mu}_d)(z_i - \hat{\mu}_d)^T.$$

Random zeroes and sparse data

If any cell in x is randomly zero, the matrix of regressors U has deficient rank and the least-squares estimate (9.10) is no longer defined. When this happens, the mean vector μ_d corresponding to the empty cell drops out of the likelihood function and becomes inestimable; the likelihood takes the same value regardless of μ_d, and the ML estimate is no longer unique.

Clearly, the unrestricted general location model will tend to be useful only when n is large relative to D, when enough observations are present in each cell to estimate all the components of μ. When the data are sparse, restricted versions of the model that contain fewer free parameters (to be discussed below) will be more appropriate.

Table 9.1. *Classification of subjects by foreign language studied and sex*

	SEX		
LAN	male	female	total
French	35	31	66
Spanish	45	32	77
German	62	52	114
Russian	9	11	20
total	151	126	277

9.2.3 Example

In Section 6.3 we examined data pertaining to the validity of the Foreign Language Attitude Scale (FLAS), a test instrument for predicting achievement in foreign language study; the raw data are reproduced in Appendix A. We will now apply the unrestricted general location model to a portion of this dataset. As shown in Table 6.5, the variables LAN and FLAS had no missing values, and SEX and HGPA were missing for only one subject each. For the moment, let us discard those two subjects to obtain an apparently complete dataset with four variables and 277 observations. The variables FLAS and HGPA are continuous, whereas LAN and SEX are categorical with four and two levels, respectively. The frequencies for the LAN by SEX classification are shown in Table 9.1. Adopting a columnwise storage order, the cell counts are

$$U^TU = \text{diag}(35, 45, 62, 9, 31, 32, 52, 11),$$

and dividing these counts by $n = 277$ yields the ML estimate

$$\hat{\pi} = (0.126, 0.162, 0.224, 0.032, 0.112, 0.116, 0.188, 0.040).$$

The sufficient statistics pertaining to HGPA and FLAS are

$$U^TZ = \begin{bmatrix} 94.45 & 2841 \\ 121.08 & 3397 \\ 170.78 & 4987 \\ 26.35 & 694 \\ 82.63 & 2759 \\ 83.12 & 2719 \\ 153.41 & 4517 \\ 20.86 & 607 \end{bmatrix}, \quad Z^TZ = \begin{bmatrix} 2199.69 & 62894.18 \\ 62894.18 & 1934421 \end{bmatrix}.$$

Dividing the rows of U^TZ by the cell counts yields the estimated matrix of means,

$$\hat{\mu} = \begin{bmatrix} 2.70 & 81.2 \\ 2.69 & 75.5 \\ 2.75 & 80.4 \\ 2.93 & 77.1 \\ 2.67 & 89.0 \\ 2.60 & 85.0 \\ 2.95 & 86.9 \\ 2.71 & 82.4 \end{bmatrix},$$

and the ML estimate of the covariance matrix is

$$\begin{aligned} \hat{\Sigma} &= n^{-1}\left(Z^TZ - Z^TU(U^TU)^{-1}U^TZ\right) \\ &= \begin{bmatrix} 0.367 & 0.411 \\ 0.411 & 176.9 \end{bmatrix}. \end{aligned}$$

9.2.4 Complete-data Bayesian inference

The factorization (9.7) which simplified the problem of ML estimation is also convenient from a Bayesian point of view: if we apply independent prior distributions to π and (μ, Σ), these parameter sets will be independent in the posterior distribution as well. For simplicity, we will apply a Dirichlet prior to the cell probabilities,

$$\pi \sim D(\alpha),$$

where $\alpha = \{\alpha_w : w \in \mathcal{W}\} = \{\alpha_d : d = 1, 2, \ldots, D\}$ is an array of user-specified hyperparameters; the complete-data posterior distribution of π is then

$$\pi \sim D(\alpha'),$$

where $\alpha' = \alpha + x$. For discussion on choosing values for the hyperparameters, see Section 7.2.5.

Inferences for μ and Σ under a noninformative prior

With regard to μ and Σ, let us first consider what happens when we apply an improper uniform prior to the elements of μ and the standard noninformative prior to the covariance matrix Σ,

$$P(\mu, \Sigma) \sim |\Sigma|^{-(\frac{p+1}{2})}$$

With a little algebra, the likelihood factor (9.8) for μ and Σ can be written in terms of the least-squares estimates,

$$L(\mu,\Sigma|W,Z) \propto |\Sigma|^{-\frac{n}{2}} \exp\Big\{-\tfrac{1}{2}\operatorname{tr}\Sigma^{-1}\hat{\varepsilon}^T\hat{\varepsilon} - \tfrac{1}{2}\operatorname{tr}\Sigma^{-1}(\mu-\hat{\mu})^T U^T U(\mu-\hat{\mu})\Big\}. \quad (9.13)$$

The diagonal form of U^TU then allows us to rewrite (9.13) as

$$L(\mu,\Sigma|Z,W) \propto |\Sigma|^{-\frac{n}{2}} \exp\Big\{-\tfrac{1}{2}\operatorname{tr}\Sigma^{-1}\hat{\varepsilon}^T\hat{\varepsilon} - \tfrac{1}{2}\sum_{d=1}^{D} x_d(\mu_d-\hat{\mu}_d)^T\Sigma^{-1}(\mu_d-\hat{\mu}_d)\Big\},$$

which is equivalent to

$$L(\mu,\Sigma|Z,W) \propto |\Sigma|^{-\left(\frac{n-p}{2}\right)} \exp\Big\{-\tfrac{1}{2}\operatorname{tr}\Sigma^{-1}\hat{\varepsilon}^T\hat{\varepsilon}\Big\} \quad (9.14)$$
$$\times \prod_{d=1}^{D} |x_d^{-1}\Sigma|^{-\frac{1}{2}} \exp\Big\{-\tfrac{1}{2}(\mu_d-\hat{\mu}_d)^T(x_d^{-1}\Sigma)^{-1}(\mu_d-\hat{\mu}_d)\Big\}.$$

Combining (9.14) with the prior (9.12) leads to

$$P(\mu,\Sigma|Z,W) \propto |\Sigma|^{-\left(\frac{n-p+q+1}{2}\right)} \exp\Big\{-\tfrac{1}{2}\operatorname{tr}\Sigma^{-1}\hat{\varepsilon}^T\hat{\varepsilon}\Big\}$$
$$\times \prod_{d=1}^{D} |x_d^{-1}\Sigma|^{-\frac{1}{2}} \exp\Big\{-\tfrac{1}{2}(\mu_d-\hat{\mu}_d)^T(x_d^{-1}\Sigma)^{-1}(\mu_d-\hat{\mu}_d)\Big\},$$

which, by inspection, is the product of independent multivariate normal densities for $\mu_1,\mu_2,\ldots,\mu_D$ given Σ and an inverted-Wishart density for Σ,

$$\mu_d \,|\, \Sigma, Y \sim N(\hat{\mu}_d, x_d^{-1}\Sigma), \quad (9.15)$$
$$\Sigma \,|\, Y \sim W^{-1}(n-D, (\hat{\varepsilon}^T\hat{\varepsilon})^{-1}). \quad (9.16)$$

For this posterior to be proper, we need $n \geq D+q$ and $x_d > 0$ for all d, structural zeroes excluded; also, the matrix $\hat{\varepsilon}^T\hat{\varepsilon}$ of residual sums of squares and cross-products should have full rank.

Informative priors

The preceding arguments can easily be extended to incorporate prior knowledge about μ and Σ. The most convenient way to specify multivariate normal distributions for $\mu_1,\mu_2,\ldots,\mu_D$ with covariance matrices proportional to Σ; prior information for Σ could then be expressed through an inverted-Wishart distribution. The resulting complete-data posterior would again be the product of independent normal distributions for $\mu_1,\mu_2,\ldots,\mu_D$ given Σ and an inverted-Wishart distribution for Σ, and the updated hyperparameters would be obtained by calculations similar to those given in Section 5.2.2.

For typical applications of the general location model, strong prior information about μ or Σ will not be available; in all our examples, we will use the noninformative prior (9.12). The use of an improper prior can lead to difficulties, especially in sparse-data situations. For many datasets, particularly if the number of cells D in the contingency table is large, we may find that portions of μ or Σ are poorly estimated or inestimable, and the posterior may be improper. When this happens, we will not attempt to stabilize the inference through informative priors for μ or Σ; rather, we will specify a more parsimonious regression model for Z given W, reducing the number of free parameters and enforcing simpler relationships between $Z_1,Z_2,\ldots,Z_q$ and $W_1,W_2,\ldots,W_p$.

9.3 Restricted models

9.3.1 Reducing the number of parameters

The unrestricted general location model tends to work well when the sample size n is appreciably larger than the total number of cells D. When this is not the case, the data may contain little or no information about certain aspects of π, μ or Σ, and it would be wise to reduce the number of free parameters. As shown by Krzanowski (1980, 1982) and Little and Schluchter (1985), the general location model is amenable to certain types of restrictions on the parameter space. Because we defined the complete-data distribution and likelihood as the product of two distinct factors, the marginal distribution of W and the conditional distribution of Z given W, we will impose restrictions on the parameter sets π and (μ,Σ) separately to keep them distinct.

Loglinear models for the cell probabilities

For the cell probabilities π, we may require them to satisfy a loglinear model

where M is a user-specified matrix. Because the contingency table is a cross-classification by $W_1, W_2, \ldots, W_p$, M will typically reflect this structure, containing 'main effects' for $W_1, W_2, \ldots, W_p$ and 'interactions' among them. If the first column of M is constant, the first element of λ (the intercept) is not a free parameter but a normalizing constant that scales π to sum to one. The total number of free parameters in this loglinear model is $\text{rank}(M) - 1$. Our fitting procedures will operate directly on the elements of π; there will be no need to explicitly create M or estimate λ unless the loglinear coefficients are of intrinsic interest.

Linear models for the within-cell means

In the unrestricted general location model, the conditional distribution of Z given W is specified by the multivariate regression

$$Z = U\mu + \epsilon, \tag{9.18}$$

where U is an $n \times D$ matrix of dummy indicators recording the cell location $1, 2, \ldots, D$ of each sample unit. The means of $Z_1, Z_2, \ldots, Z_q$ are allowed to vary freely among cells. As a result, (9.18) is equivalent to a multivariate analysis of variance (MANOVA) model for $(Z_1, Z_2, \ldots, Z_q)$ with main effects for $W_1, W_2, \ldots, W_p$ and all interactions among them. In practice, many of these interactions may be poorly estimated, and it is advantageous to eliminate them from the model.

To simplify the model, we could directly replace U by another matrix with fewer columns. For notational purposes, however, it is helpful to retain the present definition of U because of its role in the complete-data sufficient statistics. Instead, let us restrict μ to be of the form

$$\mu = A\beta \tag{9.19}$$

for some β, where A is a constant matrix of dimension $D \times r$. Each of the q columns of μ, corresponding to the variables $Z_1, Z_2, \ldots, Z_q$, is thus required to lie in the linear subspace of $\mathcal{R}^D$ spanned by the columns of A. The regression model becomes

$$Z = UA\beta + \epsilon,$$

with a reduced set of regression coefficients in β. By taking $A = I$ (the identity matrix) we obtain the unrestricted model (9.18) as a special case.

If A has full rank, then each of the $r \times q$ elements of β represents a free parameter. More generally, the number of free parameters in β is $q \times \text{rank}(A)$. If the contingency table contains no random zeroes, then all of the regression coefficients will be estimable. If the table does contain zeroes, the coefficients may still all be estimable, because estimability now depends on the rank of UA rather than U itself. To keep matters simple, let us proceed under the assumption that there are no deficiencies in the rank of A or UA,

$$\text{rank}(A) = \text{rank}(UA) = r.$$

In practice we can ensure that this is satisfied by defining A to have full rank, and then checking the rank of UA by seeing whether A^TU^TUA is invertible.

Choosing the design matrix

The design matrix A defines the regression that relates the cells of the contingency table to the means of the continuous variables. This matrix is created in the same way that one creates a design matrix for a factorial ANOVA. Thinking of the categorical variables $W_1, W_2, \ldots, W_p$ as 'factors' of the experiment, we first list all the possible combinations of levels of these factors, using the linear storage order that we adopted for our contingency table; these identify the rows of A. Then we create columns for the main effects of $W_1, W_2, \ldots, W_p$, and perhaps interactions among them, using any coding scheme that is convenient. In most applications, the first column of A will contain 1s for an intercept and the remaining columns will contain dummy codes or contrasts for the desired effects of $W_1, W_2, \ldots, W_p$ and their interactions.

For example, consider a modal with $p = 2$ categorical variables, W_1 and W_2, taking $d_1 = 2$ and $d_2 = 3$ levels, respectively, so that the contingency table has $D = 6$ cells. Let us adopt the anti-lexicographical storage order

$$(W_1, W_2) = (1,1), (2,1), (1,2), (2,2), (1,3), (2,3).$$

One possible design matrix is

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix},$$

whose columns correspond to the intercept, a main-effect contrast for W_1 and two main-effect contrasts for W_2. We may also add

contrasts for the W_1W_2 interaction by including the products of the second column with the third and fourth. If the interaction were included, the resulting model would have the same number of parameters and give the same fit as the unrestricted version (9.18).

9.3.2 *Likelihood inference for restricted models*

The two sets of restrictions that we have imposed, the loglinear restrictions on π and the linear restrictions on μ, do not interfere with each other; the joint parameter space for $\theta = (\pi, \mu, \Sigma)$ is still the product of the individual spaces for π and (μ, Σ). Therefore, the problem of maximizing the joint likelihood for θ still separates into two unrelated maximizations. The ML estimate for π may be found by conventional IPF (Section 8.3). For μ and Σ, the estimates come from the least-squares fit of the reduced regression model $Z = UA\beta + \epsilon$, which gives

$$\begin{aligned} \hat{\beta} &= (A^TU^TUA)^{-1}A^TU^TZ \\ &= (A^TT_3A)^{-1}A^TT_2, \qquad (9.20) \\ n\hat{\Sigma} &= (Z - UA\hat{\beta})^T(Z - UA\hat{\beta}) \\ &= T_1 - T_2^TA(A^TT_3A)^{-1}A^TT_2. \qquad (9.21) \end{aligned}$$

The corresponding ML estimate of μ is $\hat{\mu} = A\hat{\beta}$. For the covariance matrix, most statisticians would tend to use the unbiased estimate $n(n-r)^{-1}\hat{\Sigma}$ rather than $\hat{\Sigma}$. Notice that AT_3A is not diagonal, so in general the estimation of μ and Σ now requires the inversion of an $r \times r$ matrix.

Example: Foreign Language Attitude Scale

Returning to the example of Section 9.2.3, let us fit a reduced model to this four-variable dataset in which (a) SEX and LAN are marginally independent, and (b) the linear model for HGPA and FLAS has only main effects for SEX and LAN. Let x_{ij} denote a count in the LAN × SEX contingency table (Table 9.1) and π_{ij} the corresponding cell probability. The ML estimates of the cell probabilities for the independence model are available in closed form as $\hat{\pi}_{ij} = x_{i+}x_{+j}/n^2$, which gives

$$\hat{\pi} = (0.120, 0.152, 0.224, 0.039, 0.108, 0.126, 0.187, 0.033).$$

Using the dummy-coded design matrix

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix},$$

the least-squares regression of Z on UA yields

$$\hat{\beta} = \begin{bmatrix} 2.825 & 83.435 \\ -0.126 & 5.403 \\ -0.154 & 0.390 \\ 0.036 & 4.024 \\ -0.032 & -7.522 \end{bmatrix}, \quad \hat{\Sigma} = \begin{bmatrix} 0.372 & 0.385 \\ 0.385 & 177.3 \end{bmatrix}.$$

The corresponding ML estimate of the cell-means matrix is

$$\hat{\mu} = A\hat{\beta} = \begin{bmatrix} 2.67 & 81.3 \\ 2.64 & 76.3 \\ 2.83 & 79.9 \\ 2.79 & 75.9 \\ 2.70 & 88.8 \\ 2.67 & 83.8 \\ 2.86 & 87.5 \\ 2.82 & 83.4 \end{bmatrix}$$

We can check the plausibility of this restricted model against the unrestricted alternative by means of a likelihood-ratio test. Plugging $\hat{\pi}$, $\hat{\mu}$ and $\hat{\Sigma}$ into the formula for the complete-data log-likelihood,

$$\begin{aligned} l(\pi, \mu, \Sigma \mid Y) &= \sum_{d=1}^{D} x_d \log \pi_d - \tfrac{n}{2}\log|\Sigma| - \tfrac{1}{2}\operatorname{tr}\Sigma^{-1}T_1 \\ &\quad + \operatorname{tr}\Sigma^{-1}\mu^TT_2 - \tfrac{1}{2}\operatorname{tr}\Sigma^{-1}\mu^TT_3\mu, \qquad (9.22) \end{aligned}$$

yields a value of -1394.83. The parameter estimates from the unrestricted model (Section 9.2.3) give a slightly higher loglikelihood of -1391.86. The two models are separated by $(4-1)\times(2-1) = 3$ parameters for the marginal association between SEX and LAN, plus $3 \times 2 = 6$ coefficients for the LAN×SEX interaction in the linear model for HGPA and FLAS. The deviance statistic is $2 \times (-1391.86 + 1394.83) = 5.94$, and the corresponding p-value is $P(\chi^2_6 > 5.94) = 0.75$. The reduced model thus appears to fit the

data quite adequately. Because the complete-data likelihood factors into distinct pieces for π and (μ, Σ), we can also separate this goodness-of-fit test into two tests, one for the marginal model for LAN and SEX (3 degrees of freedom), another for the conditional model for HGPA and FLAS (6 degrees of freedom), and the two deviance statistics will add up to the overall deviance.

9.3.3 Bayesian inference

Bayesian inference for the restricted model proceeds most easily if we apply independent prior distributions to the parameter sets π and (μ, Σ), so that they remain independent in the complete-data posterior distribution. In keeping with the methods developed in the last chapter, let us apply a constrained Dirichlet prior to the elements of π, with prior density

$$P(\pi) \propto \prod_{d=1}^{D} \pi_d^{\alpha_d - 1}$$

for values of π that satisfy the loglinear constraints and $P(\pi) = 0$ elsewhere. The complete-data posterior density will then be constrained Dirichlet with updated hyperparameters $\alpha'_d = \alpha_d + x_d$. Posterior modes can be calculated using conventional IPF (Section 8.3), and simulated posterior draws of π can be obtained with Bayesian IPF (Section 8.4).

Bayesian inference for β and Σ under a noninformative prior

Bayesian inference for the standard multivariate regression model is covered in many texts on multivariate analysis; a good source is Press (1982). The likelihood function for Σ and the free coefficients β is

$$L(\beta, \Sigma | Y) \propto |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\tfrac{1}{2} \operatorname{tr} \Sigma^{-1} (Z - UA\beta)^T (Z - UA\beta) \right\}.$$

Following some algebraic manipulation, this likelihood function can be rewritten in terms of the least-squares estimates as

$$|\Sigma|^{-\frac{n}{2}} \exp\left\{ -\tfrac{1}{2} \operatorname{tr} \Sigma^{-1} \hat{\varepsilon}^T \hat{\varepsilon} - \tfrac{1}{2} (\beta - \hat{\beta})^T [\Sigma \otimes V]^{-1} (\beta - \hat{\beta}) \right\}, \quad (9.23)$$

where $\hat{\beta}$ is the matrix of estimated coefficients, $\hat{\varepsilon} = Z - UA\hat{\beta}$ is the matrix of estimated residuals and $V = (A^T U^T U A)^{-1}$. The symbol

$\otimes$ denotes the Kronecker product,

$$\Sigma \otimes V = \begin{bmatrix} \sigma_{11} V & \sigma_{12} V & \cdots & \sigma_{1q} V \\ \sigma_{12} V & \sigma_{22} V & \cdots & \sigma_{2q} V \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} V & \sigma_{q2} V & \cdots & \sigma_{qq} V \end{bmatrix}.$$

In (9.23), the columns of β and $\hat{\beta}$ have been implicitly stacked to form vectors of length rq, so that $(\beta - \hat{\beta})^T [\Sigma \otimes V]^{-1} (\beta - \hat{\beta})$ is meaningful. For some elementary properties of Kronecker products, see Mardia, Kent and Bibby (1979).

Let us first consider what happens when we apply an improper uniform prior to β and the standard Jeffreys prior to Σ,

$$P(\beta, \Sigma) \propto |\Sigma|^{-(\frac{q+1}{2})}. \quad (9.24)$$

When $A = I$, we have $\beta = \mu$, and this reduces to the noninformative prior (9.12) that we used in the unrestricted model. Combining (9.24) with the likelihood function (9.23), and using the fact that

$$|\Sigma \otimes V| = |\Sigma|^r \, |V|^q,$$

we obtain the posterior density

$$P(\beta, \Sigma | Y) \propto |\Sigma|^{-(\frac{n-r+q+1}{2})} \exp\left\{ -\tfrac{1}{2} \operatorname{tr} \Sigma^{-1} \hat{\varepsilon}^T \hat{\varepsilon} \right\} \quad (9.25)$$

$$\times \, |\Sigma \otimes V|^{-\frac{1}{2}} \exp\left\{ -\tfrac{1}{2} (\beta - \hat{\beta})^T [\Sigma \otimes V]^{-1} (\beta - \hat{\beta}) \right\}.$$

By inspection, this is the product of a multivariate normal density for β given Σ and an inverted-Wishart density for Σ,

$$\beta | \Sigma, Y \sim N(\hat{\beta}, \Sigma \otimes V), \quad (9.26)$$

$$\Sigma | Y \sim W^{-1}(n - r, (\hat{\varepsilon}^T \hat{\varepsilon})^{-1}). \quad (9.27)$$

Given Σ, the posterior distribution of each column of β is multivariate normal, centered at the corresponding column of $\hat{\beta}$ and with covariance matrix proportional to V. Marginally, the columns of β have multivariate t-distributions with $n - r$ degrees of freedom. Notice that for (9.25) to be a proper posterior density, we need $n \geq q + r$, and $\hat{\varepsilon}^T \hat{\varepsilon}$ must have full rank.

Informative priors for β and Σ

One may extend the above arguments to incorporate more substantial prior information about β and Σ. To obtain a con-

terior distribution for (β, Σ) within the normal inverted-Wishart family, however, the prior distribution must have a particular form: Σ must be inverted-Wishart, and β given Σ must be multivariate normal with a patterned covariance matrix similar to that of (9.26). The limitations of this family of priors are discussed by Press (1982). In most practical applications of the general location model, it will be difficult to quantify prior knowledge about β and Σ; all our examples will use the noninformative prior (9.24). If the posterior distribution under this prior is not proper, then we may interpret it as a sign that the model is too complex to be supported by the data, and the model should be simplified by choosing a design matrix A with fewer columns.

9.4 Algorithms for incomplete mixed data

Thus far we have reviewed the basic methods of likelihood and Bayesian inference for the parameters of the unrestricted (Section 9.2) and restricted (Section 9.3) general location models. Now we extend these methods to handle mixed datasets with arbitrary patterns of missing values. These algorithms are built from portions of the code for normal and categorical data given in Chapters 5–8. The reader who is less interested in computational details than in applications may wish to lightly skim this section to see what algorithms are available, and then proceed directly to the data examples in Section 9.5.

9.4.1 Predictive distributions

A row of the data matrix may have missing values for any or all of the variables $W_1, \ldots, W_p, Z_1, \ldots, Z_q$. Before we can derive estimation and simulation algorithms for the general location model, we must be able to characterize the joint distribution of any subset of these variables given the rest, so that we can obtain the predictive distribution of the missing data in any row of the data matrix given the observed data.

Categorical variables completely missing

Let us first consider the conditional distribution of the categorical variables given the continuous ones, which is needed when $Z_1, \ldots, Z_q$ are observed but $W_1, \ldots, W_p$ are missing. We can represent the complete data for row i by (u_i, z_i), where z_i^T is the real-

a single 1 in the cell position corresponding to the realized values of $W_1, \ldots, W_p$ and 0s elsewhere. Let E_d be the D-vector with 1 in position d and 0s elsewhere. By definition, the joint density of (u_i, z_i) under the general location model is

$$P(u_i = E_d, z_i \mid \theta) \propto \pi_d |\Sigma|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(z_i - \mu_d)^T \Sigma^{-1}(z_i - \mu_d)\right\}.$$

The conditional distribution of u_i given z_i is thus

$$P(u_i = E_d \mid z_i, \theta) = \frac{\pi_d \exp\left\{-\frac{1}{2}(z_i - \mu_d)^T \Sigma^{-1}(z_i - \mu_d)\right\}}{\sum_{d'=1}^{D} \pi_{d'} \exp\left\{-\frac{1}{2}(z_i - \mu_{d'})^T \Sigma^{-1}(z_i - \mu_{d'})\right\}}.$$

The portions of the numerator and denominator involving the quadratic term $z_i^T \Sigma^{-1} z_i$ cancel out, leading to a well-known result from classical multivariate analysis: the conditional probability that unit i belongs to cell d is

$$P(u_i = E_d \mid z_i, \theta) \propto \exp(\delta_{d,i}),$$

where $\delta_{d,i}$ denotes the value of the *linear discriminant function* of z_i with respect to μ_d,

$$\delta_{d,i} = \mu_d^T \Sigma^{-1} z_i - \tfrac{1}{2} \mu_d^T \Sigma^{-1} \mu_d + \log \pi_d. \qquad (9.28)$$

When $Z_1, \ldots, Z_q$ are observed but $W_1, \ldots, W_p$ are missing, the predictive distribution of $W_1, \ldots, W_p$ is obtained by calculating the terms $\pi_d \exp(\delta_{d,i})$ for cells $d = 1, 2, \ldots, D$ and normalizing them to sum to one.

Continuous variables partially missing

Now consider what happens if $W_1, \ldots, W_p$ and an arbitrary subset of $Z_1, \ldots, Z_q$ are missing. Denote the observed components of z_i by $z_{i(obs)}$ and the missing components by $z_{i(mis)}$. The conditional distribution of u_i given $z_{i(obs)}$ and θ is obtained by integrating both the numerator and denominator of

$$P(u_i = E_d \mid z_i, \theta) = \frac{P(u_i = E_d, z_i \mid \theta)}{P(z_i \mid \theta)}$$

over all possible values of $z_{i(mis)}$. The result is

where $\delta^*_{d,i}$ is a linear discriminant based on the reduced information in $z_{i(obs)}$ rather than z_i. This new discriminant is

$$\delta^*_{d,i} = \mu^{*T}_{d,i}\Sigma^{*-1}_i z_{i(obs)} - \tfrac{1}{2}\mu^{*T}_{d,i}\Sigma^{*-1}_i\mu^*_{d,i} + \log\pi_d, \qquad (9.30)$$

where $\mu^*_{d,i}$ and Σ^*_i denote the subvector and square submatrix of μ_d and Σ, respectively, corresponding to the observed elements of z_i. (When all continuous variables are missing, define $\delta^*_{d,i} = \log\pi_d$ so that (9.29) reduces to π_d.) Moreover, because

$$z_i \mid u_i = E_d, \theta \sim N(\mu_d, \Sigma),$$

the conditional distribution of the missing elements of z_i given $u_i = E_d$ and the observed elements of z_i is also multivariate normal; the parameters of this distribution can be obtained by applying the sweep operator to μ_d and Σ (Section 5.2). This conditional normal distribution, along with the probabilities (9.29), characterize the joint predictive distribution of $W_1, \ldots, W_p$ and the missing elements of $Z_1, \ldots, Z_q$.

Continuous and categorical variables partially missing

Finally, let us now consider the general case in which arbitrary subsets of $W_1, \ldots, W_p$ and $Z_1, \ldots, Z_q$ are missing. This differs from the case we have just examined in that the predictive distribution must now take into account any additional information in the observed members of $W_1, \ldots, W_p$. When some of these categorical variables are observed, the unit is known to lie within a particular subset of the cells of the contingency table; the cell probabilities are still of the form (9.29), but must be normalized to sum to one over this reduced set.

More specifically, let $w_{i(obs)}$ and $w_{i(mis)}$ denote the observed and missing parts, respectively, of the categorical data for unit i. Rather than indexing the cells of the contingency table by their linear positions $d = 1, 2, \ldots, D$, let us now identify them by their corresponding response patterns $w = (w_1, w_2, \ldots, w_p)$, $w_j = 1, 2, \ldots, d_j$. Let $O_i(w)$ and $\mathcal{M}_i(w)$ denote the subvectors of w corresponding to the observed and missing parts, respectively, of the categorical data for unit i. The predictive probability of falling into cell w given the observed data is now

$$P(u_i = E_w \mid w_{i(obs)}, z_{i(obs)}, \theta) = \frac{\exp(\delta^*_{w,i})}{\sum_{\mathcal{M}_i(w)} \exp(\delta^*_{w,i})} \qquad (9.31)$$

over the cells w for which $O_i(w)$ agrees with $w_{i(obs)}$, and zero for all other cells. Once again, the conditional predictive distribution of $z_{i(mis)}$ given $u_i = E_w$ is a multivariate normal whose parameters can be obtained by sweeping μ_w and Σ on the positions corresponding to $z_{i(obs)}$.

Predictive distributions and sweep

As shown by Little and Schluchter (1985), the discriminants $\delta^*_{w,i}$ and the parameters of the conditional normal distribution of $z_{i(mis)}$ can be neatly obtained by a single application of the sweep operator. Suppose we arrange the parameters of the general location model into a matrix,

$$\theta = \begin{bmatrix} \Sigma & \mu^T \\ \mu & P \end{bmatrix}, \qquad (9.32)$$

where P is a $D \times D$ matrix with elements

$$p_w = 2\log\pi_w$$

on the diagonal and zeroes elsewhere. If we sweep this θ-matrix on the positions in Σ corresponding to $z_{i(obs)}$, we obtain a transformed version of the parameter,

$$\theta^* = \begin{bmatrix} \Sigma^* & \mu^{*T} \\ \mu^* & P^* \end{bmatrix}. \qquad (9.33)$$

The diagonal element of P^* corresponding to cell w is

$$p^*_w = -\mu^{*T}_{w,i}\Sigma^{*-1}_i\mu^*_{w,i} + 2\log\pi_w,$$

which is twice the sum of the final two terms in the linear discriminant function (9.30). The coefficients of $z_{i(obs)}$ in this discriminant, $\mu^{*T}_{w,i}\Sigma^{*-1}_i$, are found in row w of μ^*, in the columns corresponding to the variables in $z_{i(obs)}$. The remaining elements of μ^* and Σ^* contain the parameters of the multivariate regression of $z_{i(mis)}$ on $z_{i(obs)}$ for all cells w. The intercepts, which vary from cell to cell, are found in μ^*; the slopes and residual covariances, which are assumed to be equal for all cells, are found in Σ^*.

Although we have depicted θ as a $(q+D) \times (q+D)$ matrix, in practice we do not actually need $(q+D)^2$ memory locations to store it. The off-diagonal elements of P^* are not really of interest, nor are they needed to reverse-sweep θ^* back to its original form. Thus we can minimize computation and memory requirements by omitting

only μ, the diagonal elements of P and the upper-triangular portion of Σ in packed storage.

9.4.2 EM for the unrestricted model

We are now ready to describe an EM algorithm for obtaining ML estimates for the unrestricted general location model (Little and Schluchter, 1985). In Section 9.2.2, we saw that the complete-data loglikelihood is a linear function of the sufficient statistics

$$T_1 = Z^T Z, \quad T_2 = U^T Z, \quad \text{and} \quad T_3 = U^T U = \text{diag}(x).$$

The ML estimates for the unrestricted model were shown to be

$$\hat{\pi} = n^{-1} x, \tag{9.34}$$

$$\hat{\mu} = T_3^{-1} T_2, \tag{9.35}$$

$$\hat{\Sigma} = n^{-1} \left(T_1 - T_2^T T_3^{-1} T_2 \right). \tag{9.36}$$

The M-step is a simple matter of calculating (9.34)–(9.36) using the expected versions of T_1, T_2 and T_3, rather than the sufficient statistics themselves. The complicated part is the E-step, where we must find the conditional expectations of T_1, T_2 and T_3 given the observed parts of the data matrix and an assumed value of θ.

The E-step

First, consider the expectation of the diagonal elements of T_3. Notice that the complete-data contingency table can be written as $x = \sum_{i=1}^{n} u_i$. The elements of u_i are Bernoulli indicators of $u_i = E_w$ for all cells w, so their expectations are just the predictive probabilities given by (9.31). Thus, the expectation of u_i can be found by the following steps. (a) Sweep the θ-matrix on positions corresponding to $z_{i(obs)}$ to obtain θ^*. (b) From $z_{i(obs)}$ and θ^*, calculate the discriminants for all cells w for which $\mathcal{O}_i(w)$ agrees with $w_{i(obs)}$. The discriminant for cell w is

$$\delta^*_{w,i} = \tfrac{1}{2} p^*_w + \sum_{j \in \mathcal{O}_i} \mu^*_{w,j} z_{ij},$$

where $\mu^*_{w,j}$ is the (w,j)th element of μ^*, and $\mathcal{O}_i$ is the subset of $\{1, 2, \ldots, q\}$ corresponding to the variables in $z_{i(obs)}$. (We have already been using $\mathcal{O}_i$ and $\mathcal{M}_i$ as operators that extract the observed and missing components of $w = (w_1, \ldots, w_p)$, and for convenience we will continue to do so; the dual usage should not create any confusion.) (c) Normalize the terms $\exp(\delta^*_{w,i})$ for these cells to obtain

the predictive probabilities

$$\pi^*_{w,i} = \frac{\exp(\delta^*_{w,i})}{\sum_{\mathcal{M}_i(w)} \exp(\delta^*_{w,i})}. \tag{9.37}$$

These predictive probabilities also play an important role in the expectation of T_2. Row w of T_2 is $\sum_{i=1}^{n} u_{w,i} x_i^T$, where $u_{w,i} = 1$ if unit i falls into cell w and $u_{w,i} = 0$ otherwise. If the observed data in $w_{i(obs)}$ indicate that unit i cannot possibly belong to cell w, then

$$E(u_{w,i} z_i \mid Y_{obs}, \theta) = 0.$$

On the other hand, if $w_{i(obs)}$ agrees with $\mathcal{O}_i(w)$, then

$$E(u_{w,i} z_i \mid Y_{obs}, \theta) = \pi^*_{w,i} z^*_{w,i}, \tag{9.38}$$

where $z^*_{w,i}$ is the predicted mean of z_i given the observed values in $z_{i(obs)}$, and given that unit i falls into cell w. The parts of $z^*_{w,i}$ corresponding to $z_{i(obs)}$ are identical to $z_{i(obs)}$, whereas the parts corresponding to $z_{i(mis)}$ are the predicted values from the multivariate regression of $z_{i(mis)}$ on $z_{i(obs)}$ within cell w,

$$z^*_{w,ij} = \begin{cases} z_{ij} & \text{if } j \in \mathcal{O}_i, \\ \mu^*_{w,j} + \sum_{k \in \mathcal{O}_i} \sigma^*_{jk} z_{ik} & \text{if } j \in \mathcal{M}_i, \end{cases}$$

where σ^*_{jk} is the (j,k)th element of Σ^*.

Finally, consider the expectation of the sums of squares and cross-products matrix,

$$T_1 = Z^T Z = \sum_{i=1}^{n} z_i z_i^T.$$

The (j,k)th element of this matrix is $\sum_{i=1}^{n} z_{ij} z_{ik}$. But notice that a single element of this sum can be written as

$$z_{ij} z_{ik} = \sum_{w} u_{w,i} z_{ij} z_{ik},$$

so the expectation of this element is

$$E(z_{ij} z_{ik} \mid Y_{obs}, \theta) = \sum_{\mathcal{M}_i(w)} \pi^*_{w,i} E(z_{ij} z_{ik} \mid Y_{obs}, \theta, u_{w,i} = 1), \tag{9.39}$$

where the sum is taken over all cells w for which $\mathcal{O}_i(w)$ agrees with $w_{i(obs)}$. The form of $E(z_{ij} z_{ik} \mid Y_{obs}, \theta, u_{w,i} = 1)$ depends on whether z_{ij} and z_{ik} are observed. If both are observed, this expectation is

simply $x_{ij}\, x_{ik}$. If z_{ij} is observed but z_{ik} is missing, the expectation is $z_{ij}\, z^*_{w,jk}$. Finally, if both are missing, the expectation becomes $z^*_{w,ij}\, z^*_{w,ik} + \sigma^*_{jk}$.

Organizing the computations

To carry out the E-step, we must cycle through the units $i = 1, 2, \ldots, n$ in the dataset, sweeping θ on the positions corresponding to $z_{i(obs)}$ and summing the contributions (9.37), (9.38) and (9.39) of unit i to the expectations of the sufficient statistics. The number of forward and reverse-sweeps can be reduced by grouping together rows of the data matrix having the same pattern of missingness for $Z_1, \ldots, Z_q$, because the same version of θ^* can then be used for all units in the pattern. The expected sufficient statistics can be accumulated into a workspace of the same size and shape as θ,

$$T = \begin{bmatrix} T_1 & T_2^T \\ T_2 & T_3 \end{bmatrix}.$$

Once the E-step is complete, the M-step proceeds by applying (9.34)–(9.36) to T, which gives the updated estimate of θ.

Evaluating the observed-data loglikelihood

One can show that the contribution of observation i to the observed-data loglikelihood is

$$-\tfrac{1}{2}\log|\Sigma^*_i| + \log\left\{\sum_w \exp\left(\delta^*_{w,i} - \tfrac{1}{2}\, z^T_{i(obs)}\Sigma^{*-1}_i z_{i(obs)}\right)\right\},$$

where the sum is taken over all cells w for which $C_i(w)$ agrees with $w_{i(obs)}$. The procedure for evaluating the observed-data loglikelihood at any particular value of θ is very similar to the E-step. In addition to the linear discriminant $\delta^*_{w,i}$, we need to evaluate the quadratic term

$$z^T_{i(obs)}\Sigma^{*-1}_i z_{i(obs)}$$

and the determinant of Σ^{*-1}_i. The latter can be obtained along with θ^* as an immediate byproduct of sweep (Section 5.2.4). To calculate the former, note that $-\Sigma^{*-1}_i$ is contained in the rows and columns of Σ^* corresponding to the variables in $z_{i(obs)}$.

9.4.3 Data augmentation

With fairly minor modifications, the EM algorithm described above can be converted to data augmentation, enabling us to simulate posterior draws of θ or multiple imputations of Y_{mis}. For the I-step, we must create a random draw of (T_1, T_2, T_3) from its predictive distribution given the observed data and an assumed value for θ. Just as in the E-step, we cycle through the units $i = 1, 2, \ldots, n$, sweeping θ to obtain the parameters of the predictive distribution of the missing variables given the observed variables; we then draw the missing data for unit i from their predictive distribution, and accumulate the resulting complete-data sufficient statistics into T_1, T_2 and T_3. Once the I-step is complete, the P-step proceeds by drawing a new value of θ from its posterior given T_1, T_2 and T_3. Details of these steps are given below.

The I-step

It is convenient to draw the missing data for unit i in two stages: first by drawing u_i, which indicates the cell to which unit i belongs, and then by drawing $z_{i(mis)}$ given u_i. The predictive distribution of u_i is that of a single multinomial trial over the cells w for which $C_i(w)$ agrees with $w_{i(obs)}$; the cell probabilities are given by (9.37). A simple way to simulate this multinomial trial is by table sampling: cycle through the cells, summing up their probabilities, and assign the unit to the first cell for which the cumulative probability exceeds the value of a $U(0,1)$ random variate. Pseudocode for a similar table-sampling algorithm appears in Figure 7.4. When a unit is assigned to cell w, its contribution to T_3 is reflected by adding 1 to the wth diagonal element.

After assigning unit i to cell w, we may then draw the missing continuous variables in $z_{i(mis)}$ according to their multivariate regression on $z_{i(obs)}$. The regression prediction for an element of $z_{i(mis)}$ is

$$z^*_{w,ij} = \mu^*_{w,j} + \sum_{k \in O_i} \sigma^*_{jk}\, z_{ik}.$$

To these predictions, we must add simulated residuals drawn from a multivariate normal distribution. The residual covariances are found in Σ^*, in the rows and columns corresponding to $z_{i(mis)}$. To draw the residuals, we will need to extract the appropriate submatrix from Σ^* and calculate its Cholesky factor (Section 5.4.1). Adding the simulated residuals to the z^* ... produces a simulated

draw of $z_{i(mis)}$. The contribution of the completed version of z_i to the sufficient statistics is then reflected by adding z_i into the wth row of T_2, and adding $z_i z_i^T$ into the matrix T_1.

The P-step

In Section 9.2.4, we showed that under the improper prior distribution

$$P(\pi, \mu, \Sigma) \propto \left(\prod_w \pi_w^{\alpha_w - 1} \right) |\Sigma|^{-\left(\frac{q+1}{2}\right)},$$

the complete-data posterior is

$$\pi \mid Y \sim D(\alpha + x), \tag{9.40}$$

$$\Sigma \mid \pi, Y \sim W^{-1}(n - D, (\hat{\epsilon}^T \hat{\epsilon})^{-1}), \tag{9.41}$$

$$\mu_w \mid \pi, \Sigma, Y \sim N(\hat{\mu}_w, x_w^{-1} \Sigma), \tag{9.42}$$

where $\alpha = \{\alpha_w\}$ is an array of user-specified hyperparameters. The P-step is simply a matter of drawing from these distributions in turn, given the simulated values of T_1, T_2 and T_3 from the I-step. This can be done as follows.

1. For each cell w, draw the probability π_w from a standard gamma distribution with shape parameter $x_w + \alpha_w$, where x_w is the wth diagonal element of T_3, and normalize the π_w to sum to one.

2. Draw an upper-triangular matrix B whose elements are independently distributed as

$$b_{jj} \sim \sqrt{\chi^2_{n-D-j+1}}, \quad j = 1, \ldots, q,$$

$$b_{jk} \sim N(0, 1), \quad j < k,$$

and take $\Sigma = M^T M$, where $M = (B^T)^{-1} C$ and C is the upper-triangular Cholesky factor of

$$\hat{\epsilon}^T \hat{\epsilon} = T_1 - T_2^T T_3^{-1} T_2.$$

3. Calculate $\hat{\mu} = T_3^{-1} T_2$ and take $\mu = \hat{\mu} + T_3^{-1/2} H M$, where H is a $D \times q$ matrix of independent $N(0, 1)$ random variates, and $T_3^{-1/2}$ is the matrix with elements $x_w^{-1/2}$ on the diagonal and zeroes elsewhere.

9.4.4 Algorithms for restricted models

An ECM algorithm

Little and Schluchter (1985) discussed an EM algorithm for ML estimation under restricted versions of the general location model. The E-step is identical to that described above for the unrestricted model, because the expectations of T_1, T_2 and T_3 have the same form regardless of where $\theta = (\pi, \mu, \Sigma)$ lies in the parameter space. The only difference is found in the M-step, which is now a constrained maximization subject to loglinear restrictions on π and linear restrictions on μ. As discussed in Section 9.3, the constrained maxima for π and μ may be found by conventional IPF and least squares, respectively.

In the same article, Little and Schluchter also conjectured that the full maximization of the likelihood for π in each M-step, which may require many IPF cycles, could be replaced by a single IPF cycle, thus avoiding undesirable nested iterations. The resulting algorithm would no longer be EM, but it would have the same essential property that the observed-data loglikelihood would be non-decreasing. Their conjecture turned out to be correct. This algorithm is a special case of ECM, exhibiting the same reliable convergence properties as EM; see Sections 3.2.5 and 8.5.1 for further details and references. A single cycle of ECM for the restricted general location model proceeds as follows.

1. *E-step:* Given the current estimate $\theta^{(t)} = (\pi^{(t)}, \mu^{(t)}, \Sigma^{(t)})$, calculate the expectations of T_1, T_2 and T_3 as described in Section 9.4.2.

2. *CM-step:* Using the expected value of x (the diagonal elements of T_3), perform a single cycle of conventional IPF from the starting value $\pi^{(t)}$ to obtain $\pi^{(t+1)}$. Then calculate $\beta^{(t+1)}$ and $\Sigma^{(t+1)}$ as in (9.20)–(9.21) using the expected values of T_1, T_2 and T_3, and take $\mu^{(t+1)} = A\beta^{(t+1)}$.

Data augmentation-Bayesian IPF

In a similar fashion, the data augmentation algorithm for the unrestricted model can be adapted to restricted models. The I-step remains the same; only the P-step must be changed to accommodate the restrictions on the parameter space.

Under the family of prior distributions discussed in Section 9.3.3, the complete-data posterior distribution for π in

let, and the complete-data posterior for (β, Σ) is

$$\Sigma \mid Y \sim W^{-1}(n-r, (\hat{\epsilon}^T\hat{\epsilon})^{-1}), \quad (9.43)$$

$$\beta \mid \Sigma, Y \sim N(\hat{\beta}, \Sigma \otimes V), \quad (9.44)$$

where $V = (A^T U^T U A)^{-1}$. Random draws from the constrained Dirichlet can be simulated by Bayesian IPF (Section 8.4), and drawing from (9.43) is straightforward. In many applications the dimension of β can be quite large, but simulating draws from (9.44) is not difficult if we exploit the patterned covariance structure. Let G and H denote the upper-triangular Cholesky factors of Σ and V, respectively, so that $\Sigma = G^T G$ and $V = H^T H$. Using elementary properties of Kronecker products,

$$\begin{aligned} \Sigma \otimes V &= (G^T G) \otimes (H^T H) \\ &= (G^T \otimes H^T)(G \otimes H) \\ &= (G \otimes H)^T (G \otimes H), \end{aligned}$$

and thus $G \otimes H$ is an upper-triangular square root for $\Sigma \otimes V$. Therefore, to simulate a multivariate normal random vector with covariance matrix $\Sigma \otimes V$, we may simply premultiply a vector of standard normal variates by $(G \otimes H)^T$.

A data augmentation-Bayesian IPF (DABIPF) algorithm for the restricted general location model proceeds as follows.

1. *I-step:* Given the current values of the parameters $\pi^{(t)}$, $\mu^{(t)} = A\beta^{(t)}$ and $\Sigma^{(t)}$, draw the missing data from their predictive distribution as described in Section 9.4.3, and accumulate the simulated values of the sufficient statistics T_1, T_2 and T_3.
2. *Bayesian IPF:* Using the simulated value of x (the diagonal elements of T_3), perform a single cycle of Bayesian IPF from the starting value $\pi^{(t)}$ to obtain $\pi^{(t+1)}$.
3. *P-step for* Σ: Draw an upper-triangular matrix B whose elements are independently distributed as

$$b_{jj} \sim \sqrt{\chi^2_{n-r-j+1}}, \quad j = 1, \ldots, q,$$

$$b_{jk} \sim N(0,1), \quad j < k,$$

and take $\Sigma^{(t+1)} = M^T M$, where $M = (B^T)^{-1} C$ and C is the upper-triangular Cholesky factor of

$$\hat{\epsilon}^T\hat{\epsilon} = T_1 - T_2^T A (A^T T_3 A)^{-1} A^T T_2.$$

4. *P-step for* β: Draw $\beta^{(t+1)}$ from a multivariate normal distribution with mean $\hat{\beta} = (A^T T_3 A)^{-1} A^T T_2$ and covariance matrix $\Sigma^{(t+1)} \otimes V$, where $V = (A^T T_3 A)^{-1}$. This can be done in the following manner. Let β_j and $\hat{\beta}_j$ denote the jth columns of $\beta^{(t+1)}$ and $\hat{\beta}$, respectively. Calculate $G = \text{Chol}(\Sigma^{(t+1)})$ and $H = \text{Chol}(V)$, and take

$$\begin{aligned} \beta_1 &= \hat{\beta}_1 + g_{11} H^T \kappa_1, \\ \beta_2 &= \hat{\beta}_2 + g_{21} H^T \kappa_1 + g_{22} H^T \kappa_2, \\ &\vdots \\ \beta_q &= \hat{\beta}_q + g_{q1} H^T \kappa_1 + g_{q2} H^T \kappa_2 + \cdots + g_{qq} H^T \kappa_q, \end{aligned}$$

where g_{ij} is the (i,j)th element of G, and where $\kappa_1, \kappa_2, \ldots, \kappa_q$ are vectors of independent $N(0,1)$ random variates of length r.

This DABIPF algorithm is not true data augmentation, but a hybrid that substitutes a single cycle of Bayesian IPF for the full simulation of π in the P-step.

9.5 Data examples

9.5.1 St. Louis Risk Research Project

Little and Schluchter (1985) presented data from the St. Louis Risk Research Project, an observational study to assess the effects of parental psychological disorders on various aspects of child development. In a preliminary cross-sectional study, data were collected on 69 families having two children each. The families were classified into three risk groups for parental psychological disorders. The children were classified into two groups according to the number of adverse psychiatric symptoms they exhibited. Standardized reading and verbal comprehension scores were also collected for the children. Each family is thus described by three continuous and four categorical variables:

Variable	*Levels*	*Code*
Parental risk group	1=low, 2=moderate, 3=high	G
Symptoms, child 1	1=low, 2=high	D_1
Symptoms, child 2	1=low, 2=high	D_2
Reading score, child 1	continuous	R_1
Verbal score, child 1	continuous	V_1
Reading score, child 2	continuous	R_2
Verbal score, child 2	continuous	V_2

Data from this preliminary study are displayed in Table 9.2. Missing values occur on all variables except G. Only twelve families have values recorded for all seven variables.

The unrestricted model

The unrestricted general location model for this dataset has 69 free parameters: 11 for the $3 \times 2 \times 2$ contingency table that cross-classifies families by G, D_1 and D_2, 48 for the within-cell means of R_1, V_1, R_2 and V_2, and 10 for the within-cell covariance matrix. As pointed out by Little and Schluchter (1985), all of the parameters of this model are technically estimable. There are no zero counts in the table for the 29 families that can be fully classified on G, D_1 and D_2. The only family known to belong to the $G = 2$, $D_1 = 2$, $D_2 = 1$ cell has missing values for V_1, R_2 and V_2; there are three other partially classified families that can possibly belong to this cell, however, and two of these families have all their continuous variables recorded. Similarly, the only family known to have $G = 1$, $D_1 = 2$, $D_2 = 2$ has a missing value for R_2, but there are two other partially classified families for whom R_2 is known who may belong to this cell. These partially classified families contribute 'fractional observations' of the continuous variables to certain cells. With respect to the means of these cells, the observed-data likelihood is not flat, but some of the means may be estimated with precision equivalent to sample sizes of less than one.

Using the EM algorithm described in Section 9.4.2, Little and Schluchter (1985) discovered that the observed-data likelihood for this example is multimodal. They found that EM converges to different parameter estimates from different starting values, and the loglikelihood values at these estimates are not identical. The data augmentation algorithm of Section 9.4.3, when used in conjunction with EM, provides an additional tool to help us explore the observed-data likelihood. Starting at a mode, we ran several hundred iterations of data augmentation, and used the final simulated value of the parameter as a new starting value for EM. By repeating this process, we were quickly able to identify ten distinct modes, and would have undoubtedly found more had we continued further. The unusual shape of the observed-data loglikelihood suggests that some of the parameters of the unrestricted model are very poorly estimated. This is not surprising, given that we are trying to estimate 69 parameters from only 69 incomplete observations. Time-series plots of some parameters across the iterations of

Table 9.2. *Data from the St. Louis Risk Research Project*

Low risk ($G = 1$)						Moderate risk ($G = 2$)						High risk ($G = 3$)					
R_1	V_1	D_1	R_2	V_2	D_2	R_1	V_1	D_1	R_2	V_2	D_2	R_1	V_1	D_1	R_2	V_2	D_2
110	—	—	—	150	1	88	85	2	76	78	—	98	110	—	112	103	2
118	165	1	—	130	2	—	98	—	114	133	—	127	138	1	92	118	1
116	145	2	114	125	—	108	103	2	90	100	2	113	—	—	—	—	—
—	—	—	126	—	—	113	—	2	95	115	2	107	93	—	92	75	2
118	140	1	118	123	—	—	65	—	97	68	2	—	—	1	101	—	2
—	120	—	105	139	—	118	—	2	—	—	2	114	—	2	87	98	2
—	—	—	96	113	—	92	—	2	110	96	1	56	58	2	88	105	1
138	169	1	130	140	—	90	—	1	—	88	2	96	95	1	87	100	2
115	158	1	—	—	—	98	123	—	112	115	—	126	185	2	118	133	2
—	145	2	139	185	2	113	110	—	114	120	—	—	—	—	130	195	—
125	138	1	105	133	1	102	130	—	130	135	—	—	—	—	116	—	2
120	160	—	109	150	—	89	113	2	91	75	2	84	45	2	82	53	2
—	133	—	98	108	—	90	80	2	109	88	2	128	—	2	121	—	2
—	—	—	115	140	2	—	—	1	88	78	1	—	120	1	108	118	2
115	158	2	—	135	1	75	63	1	—	—	2	105	138	1	100	140	1
112	115	2	93	140	—	93	—	—	115	138	2	88	118	—	74	75	—
133	168	1	126	158	2	—	170	—	115	129	2	—	—	—	84	103	—
118	180	1	116	148	—	123	130	—	104	123	2	—	—	—	—	—	—
128	—	1	110	155	1	114	—	1	113	—	2	—	—	—	—	—	—
100	—	1	101	120	1	—	—	2	—	103	2	—	—	—	—	—	—
118	138	1	—	110	1	113	—	2	82	—	2	—	—	—	—	—	—
103	108	—	—	—	—	117	—	—	114	—	1	—	—	—	—	—	—
121	155	1	—	100	2	122	—	—	—	—	2	—	—	—	—	—	—
—	—	—	—	—	2	105	—	—	—	—	1	—	—	—	—	—	—
—	—	—	104	118	1	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	87	85	1	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	63	—	—	—	—	—	—	—	—	—	—	—	—	—

Source: Little and Schluchter (1985)

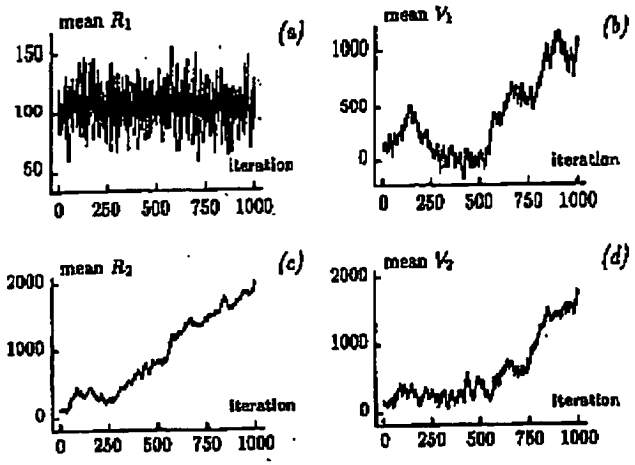


Figure 9.2. *Time-series plots of the conditional means of R_1, V_1, R_2 and V_2 given ($G = 2, D_1 = 2, D_2 = 1$) for 1000 iterations of data augmentation under the unrestricted general location model.*

data augmentation show erratic behavior. Plots of the simulated means of the four continuous variables within the $G = 2$, $D_1 = 2$, $D_2 = 1$ cell are shown in Figure 9.2. The means for V_1, R_2 and V_2 are highly unstable, wandering well outside the plausible range of reading and verbal scores. The use of the unrestricted model is not recommended for this dataset, as it is clearly overparameterized.

Restricted models.

Because the ultimate purpose of the St. Louis Risk study was to examine the relationship of parental psychological disorders on child development, we now examine two restricted models that focus attention on the effects of greatest interest, namely, the associations between parental risk G and the child development variables D_1, R_1, V_1, D_2, R_2 and V_2.

The first model, which will be called the 'null model', allows the six development variables to be interrelated, but assumes that they are collectively independent of G. The loglinear model for the categorical variables is (G, D_1D_2). The design matrix specifying the regression of the four continuous variables on the categorical ones is shown in Table 9.3 (a); it includes an intercept, main effects for D_1 and D_2 and the D_1D_2 interaction. This model fits 5 free

Table 9.3. *Design matrix for the null model, and the linear contrast for G included in the alternative model*

Cell			*(a) Design matrix*				*(b) G effect*
G	D_1	D_2	int.	D_1	D_2	D_1D_2	linear
1	1	1	1	−1	−1	1	−1
2	1	1	1	−1	−1	1	0
3	1	1	1	−1	−1	1	1
1	2	1	1	1	−1	−1	−1
2	2	1	1	1	−1	−1	0
3	2	1	1	1	−1	−1	1
1	1	2	1	−1	1	−1	−1
2	1	2	1	−1	1	−1	0
3	1	2	1	−1	1	−1	1
1	2	2	1	1	1	1	−1
2	2	2	1	1	1	1	0
3	2	2	1	1	1	1	1

parameters to the contingency table, 16 regression coefficients and 10 covariances for a total of 31 free parameters.

The second model, which we call the 'alternative model', adds simple associations between G and each of the six development variables. The loglinear model is now (GD_1, GD_2, D_1D_2), and the association between G and the continuous variables is specified by adding columns to the design matrix for G. To conserve parameters, we add only a single column for a linear contrast, as shown in Table 9.3 (b). The alternative model has 9 parameters for the contingency table, 20 regression coefficients and 10 covariances for a total of 39 parameters.

ML estimates under these two models were computed using the ECM algorithm of Section 9.4.4. As with the unrestricted model, the observed-data loglikelihood functions are not unimodal; we found two modes under the null model and two modes under the alternative. The likelihood-ratio test statistic based on the two major modes is 21.9 with 8 degrees of freedom. It appears that the alternative model may fit the data substantially better than the null model, but we cannot assign an accurate p-value to this difference due to the unusual shape of the likelihood function.

Adopting a Bayesian approach, however, we can demonstrate rather conclusively that G is indeed related to each of the six development variables. Using the DABIPF algorithm, we simulated

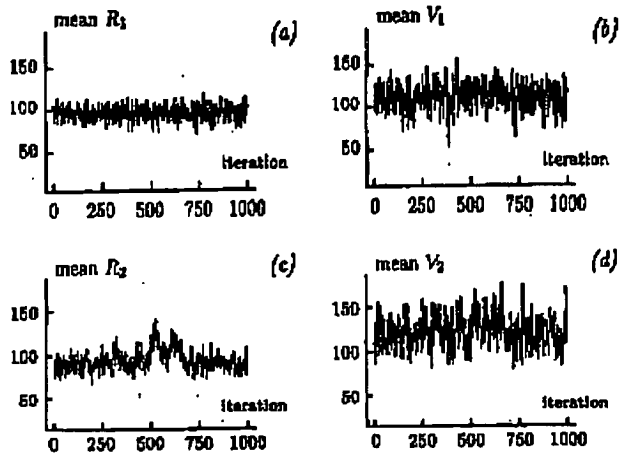


Figure 9.3. *Time-series plots of the conditional means of R_1, V_1, R_2 and V_2 given $(G = 2, D_1 = 2, D_2 = 1)$ for 1000 iterations of DABIPF under the alternative model.*

5000 correlated draws from the observed-data posterior under the alternative model and stored the values of parameters of interest. Time-series plots of the parameters, shown in Figure 9.3, did not exhibit the same instability found in plots for the unrestricted model, so the algorithm appears to be converging reliably. By examining the simulated values of the parameters pertaining to the associations between G and the other variables, we may proceed to make Bayesian inferences about these parameters directly without appealing to large-sample approximations.

Risk and adverse psychological symptoms

Let π_{ijk} denote the marginal probability of the event $G = i$, $D_1 = j$, $D_2 = k$. The association between G and D_1 can be described by two odds ratios, say

$$\omega_1 = \frac{\pi_{11k}\pi_{22k}}{\pi_{21k}\pi_{12k}}, \quad \omega_2 = \frac{\pi_{21k}\pi_{32k}}{\pi_{31k}\pi_{22k}}.$$

These express the increase in odds of adverse symptoms in the first child as we move from low to moderate risk, and from moderate to high risk, respectively. Notice that these odds ratios do not depend on k; they are identical for $k = 1$ and $k = 2$ because the loglinear model omits the three-way association GD_1D_2. Similarly,

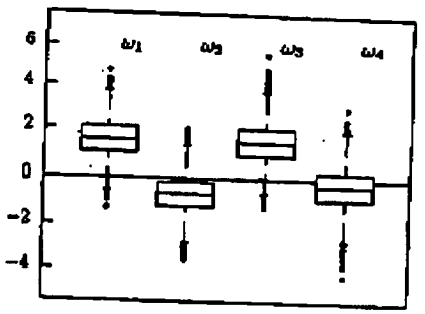


Figure 9.4. *Boxplots of simulated log-odds ratios from 5000 iterations of DABIPF under the alternative model.*

Table 9.4. *Simulated posterior percentiles and p-values for odds ratios*

	percentile					
	2.5	25	50	75	97.5	p
ω_1	1.09	2.79	4.68	8.11	22.97	0.04
ω_2	0.11	0.29	0.49	0.84	2.27	0.37
ω_3	0.97	2.66	4.50	7.90	23.67	0.05
ω_4	0.12	0.41	0.74	1.35	4.31	0.75

the association between G and D_2 can be described by

$$\omega_3 = \frac{\pi_{1j1}\pi_{2j2}}{\pi_{2j1}\pi_{1j2}}, \quad \omega_4 = \frac{\pi_{2j1}\pi_{3j2}}{\pi_{3j1}\pi_{2j2}},$$

which express the increase in odds of adverse symptoms in the second child as we move from low to moderate and from moderate to high risk.

Boxplots of the logarithms of the four odds ratios from 5000 cycles of DABIPF are shown in Figure 9.4. The logs of ω_1 and ω_3 are nearly all positive, providing strong evidence that children in moderate-risk families $(G = 2)$ have higher rates of adverse symptoms that children in low-risk families $(G = 1)$. The logs of ω_2 and ω_4, however, lie on both sides of zero; there is no evidence that the adverse-symptom rates differ for children in moderate- $(G = 2)$ and high-risk $(G = 3)$ families. Simulated percentiles of

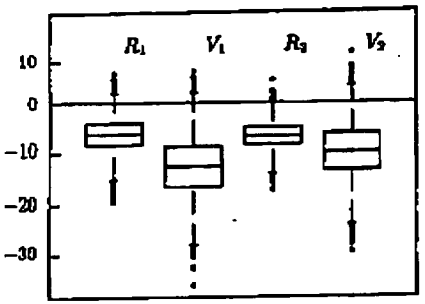


Figure 9.5. *Boxplots of simulated regression coefficients from 5000 iterations of DABIPF under the alternative model.*

Table 9.5. *Simulated posterior percentiles and p-values for regression coefficients*

	percentile					
	2.5	25	50	75	97.5	p
R_1	−12.75	−8.51	−6.38	−4.21	0.10	0.05
V_1	−23.49	−16.50	−12.62	−8.85	−1.29	0.03
R_2	−11.94	−8.64	−6.92	−5.12	−1.55	0.01
V_2	−20.37	−13.42	−10.02	−6.42	0.17	0.05

the posterior distributions of the ω_i are shown in Table 9.4, along with Bayesian p-values for testing each null hypothesis $\omega_i = 1$ against the two-sided alternative $\omega_i \neq 1$. Based on the posterior medians, we estimate that children in moderate-risk families are about 4.5 times as likely (on the odds scale) to display adverse symptoms than children in low-risk families.

Risk and comprehension scores

The association between risk and comprehension is summarized by the coefficients of the linear term for G in the regression model for R_1, V_1,R_2 and V_2. Boxplots of the simulated regression coefficients from DABIPF are displayed in Figure 9.5. For each coefficient, the majority of the simulated values lie well below zero, providing evidence that increasing risk is associated with decreasing reading and

verbal comprehension. Simulated posterior percentiles for the four coefficients are given in Table 9.5, along with a two-tailed Bayesian p-value for testing the null hypothesis that each coefficient is zero. All four effects are 'statistically significant.' From the medians, we estimate that increasing risk by one category (low to moderate or moderate to high) is associated with a drop of 6–7 points in reading comprehension and 10–13 points in verbal comprehension for each child.

9.5.2 Foreign Language Attitude Scale

In Section 6.3, we examined data pertaining to the Foreign Language Attitude Scale (FLAS), an instrument designed to predict achievement in the study of foreign languages. Of the twelve variables in the dataset, five are categorical and seven are continuous. The analyses in Chapter 6 relied on multiple imputations created under a multivariate normal model. Prior to imputation, we recoded some of the categorical variables to make the normal model appear more reasonable. In the process of recoding, however, some potentially useful detail was lost. For example, the final grade variable GRD was collapsed from five categories to only two. Now, using the general location model, we will re-impute the missing data without altering any of the categorical variables.

The imputation model

For imputation purposes, we fitted a restricted version of the general location model to the twelve variables listed in Table 6.5. The categorical variables LAN, AGE, PRI, SEX and GRD define a five-dimensional contingency table with $4 \times 5 \times 5 \times 2 \times 5 = 1000$ cells. This table was described by a loglinear model with all main effects and two-variable associations. The seven continuous variables were then described by a regression with main effects for each categorical variable. The design matrix, which had 1000 rows and eight columns, included a constant term for the intercept, three dummy indicators for LAN, a dummy indicator for SEX and linear contrasts for AGE, PRI and GRD. The coding scheme for the design matrix is shown in Table 9.6.

Like the multivariate normal distribution, this model allows simple associations between any two variables. Imputations generated under the model will preserve simple marginal and conditional associations, but higher-order effects such as interactions will not be

Table 9.6. *Columns of design matrix in imputation model, foreign language achievement study data*

Variable	*Description*
INT	constant term for intercept (1)
LAN_2	Spanish indicator (1=Spanish, 0=other language)
LAN_3	German indicator (1=German, 0=other language)
LAN_4	Russian indicator (1=Russian, 0=other language)
AGE_L	linear contrast for age (−2=less than 20, −1=20–21, 0=22–23, 1=24–25, 2=26+)
PRI_L	linear contrast for prior courses (−2=none, −1=1, 0=2, 1=3, 2=4+)
SEX_2	female indicator (1=female, 0=male)
GRD_L	linear contrast for grade (−2=F, −1=D, 0=C, 1=B, 2=A)

reflected in the imputed values. If the post-imputation analyses involve only simple associations (e.g. regressions with main effects but no interactions) then this imputation model may be expected to perform well. More elaborate analyses involving interactions, however, would require a more elaborate imputation model.

Prior distributions

Recall from Section 6.3 that certain parameters of the normal model were inestimable, because values of GRD were missing for all students enrolled in Russian (LAN=4). In the new imputation model, some aspects of the association between GRD and LAN are again inestimable for the same reason. Furthermore, the sparseness of the contingency table (recall that there are 1000 cells but only $n = 279$ observations) results in ML estimates on the boundary of the parameter space. These difficulties can be addressed by specifying a proper prior distribution for the cell probabilities.

In previous examples involving sparse tables, we applied flattening priors, Dirichlet or constrained Dirichlet distributions with hyperparameters set to a small positive constant. Flattening priors smooth the estimated cell probabilities toward a uniform table. This type of smoothing may be undesirable in this application, because some of the categorical variables (AGE and GRD, in particular) have categories that are quite rare; flattening priors

could distort the marginal distributions for these variables, leading to an over-representation of rare categories in the imputed values. Another possibility is a data-dependent prior that smooths the estimates toward a table of mutual independence among the variables, but leaves the marginal distribution of each variable unchanged (Section 7.2.5). To generate multiple imputations, we ran DABIPF under two different priors: (a) a data-dependent prior of this type, with hyperparameters scaled to add to 50; and (b) the Jeffreys prior with all hyperparameters equal to 1/2. The latter may arguably result in overamoothing; we include it primarily to assess the sensitivity of our results to the choice of prior.

Generating the imputations

Under each prior, we generated $m = 10$ imputations by running a single chain of DABIPF, allowing 250 cycles between imputations. To obtain a starting value of θ, we first ran the ECM algorithm, setting hyperparameters to 1.05 to ensure a mode in the interior of the parameter space. The continuous variables were modeled and imputed on their original scales without transformation. The imputed values for these variables hardly ever strayed outside their natural ranges. For example, only two of the $10 \times 34 = 340$ values of CGPA imputed in the first DABIPF run fell above the maximum of 4.0. Because these 'impossible' imputations occurred so rarely, we simply allowed them to remain in the imputed data rather than editing or re-drawing them.

A proportional-odds model

In keeping with the purpose of this study, a model was fitted to predict final grade GRD from the other eleven variables. Because GRD is an ordinal scale (0=F, 1=D, 2=C, 3=B, 4=A), we used a logistic model for ordinal responses known as the proportional-odds model (McCullagh, 1980; Agresti, 1990). For any subject i, let π_{ij} denote the probability of the event GRD $\geq j$, and let x_i be a vector of covariates. The proportional-odds model is

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \alpha_j + x_i^T \beta, \quad j = 1, 2, 3, 4.$$

In other words, the log-odds of falling above each of the four GRD cut-points are simultaneously modeled as parallel linear functions with common slopes β and intercepts $\alpha_1 \geq \alpha_2 \geq \alpha_3 \geq \alpha_4$. Routines for maximum-likelihood estimation in the proportional odds

Table 9.7. *Estimates, standard errors, p-values and percent missing information for coefficients in the proportional-odds model, from m = 10 multiple imputations under (a) data-dependent and (b) Jeffreys priors*

	(a) Data-dependent				(b) Jeffreys			
variable	est.	SE	p	mis.	est.	SE	p	mis.
INT_1	−6.69	2.07	.00	35	−8.38	1.86	.00	22
INT_2	−9.00	2.13	.00	42	−10.2	1.87	.00	20
INT_3	−11.3	2.19	.00	42	−12.3	1.92	.00	19
INT_4	−13.7	2.23	.00	38	−14.6	2.00	.00	19
LAN_2	−.203	.399	.61	16	−.113	.398	.77	16
LAN_3	.684	.378	.07	15	.708	.371	.06	13
LAN_4	−.857	1.08	.44	76	−2.70	1.06	.02	76
AGE_L	.381	.201	.06	25	.213	.266	.43	65
PRI_L	.344	.109	.00	32	.371	.116	.00	40
SEX_2	.338	.352	.34	22	.318	.368	.39	30
FLAS × 10	.452	.128	.00	36	.456	.132	.00	38
MLAT	.104	.046	.03	60	.130	.041	.00	46
SATV × 100	−.290	.266	.28	31	−.223	.296	.45	48
SATM × 100	.029	.207	.89	17	.165	.197	.40	10
ENG × 10	−.027	.195	.89	44	−.139	.185	.46	37
HGPA	2.21	.348	.00	29	2.23	.317	.00	16
CGPA	.912	.436	.04	31	.752	.422	.08	28

model are available in several popular statistical software packages, including SAS (SAS Institute Inc., 1990) and BMDP (BMDP Statistical Software, Inc., 1992).

The covariates in our proportional-odds model included all seven of the continuous variables in the dataset. In addition, we included three dummy indicators for LAN, a dummy indicator for SEX and linear contrasts for AGE and PRI, coded as shown in Table 9.6. For each imputed dataset, we calculated ML estimates using software developed by Harrell (1990) for the statistical system S (Becker, Chambers and Wilks, 1988). The estimates, along with standard errors based on score statistics, were then combined using Rubin's rules for scalar estimands (Section 4.3.2). Estimated coefficients and standard errors are displayed in Table 9.7, along with percent missing information and two-tailed p-values for testing the null hypothesis that each coefficient is zero.

Results using a data-dependent prior, shown in Table 9.7 (a), are fairly consistent with our findings in Section 6.3 where we fitted a

simple logit model to the dichotomized version of GRD. The only substantial difference is that under the proportional-odds model, PRI has a significant effect on GRD but SEX does not; under the dichotomous model, SEX had a significant effect but PRI did not. Results under the Jeffreys prior, shown in Table 9.7 (b), are similar to those from the data-dependent prior, with the following two exceptions: first, the linear effect of AGE is no longer significant; second, the coefficient of the dummy indicator LAN_4 is now highly significant. The latter is rather curious, because we know that the data provide essentially no information about the effect of LAN_4 on GRD given the other variables. This 'statistically significant' relationship appears to be a figment of the Jeffreys prior, which smooths the data quite heavily. The high fraction of missing information for this coefficient, along with its sensitivity to the choice of prior, should alert us to use extreme caution when trying to make any inferences regarding grades for the LAN = 4 group.

Partial correlation coefficients

Apart from determining which predictors are significantly related to GRD, it is also useful to consider the practical importance of the estimated effects. In many areas of social science, associations are expressed and compared in terms of simple or partial correlation coefficients. In linear regression, a partial correlation measures the expected change in the response variable (expressed in standard units) associated with a one-unit increase in a predictor (also in standard units) when all other predictors are held constant. A squared partial correlation measures the proportion of variance in the response variable 'explained by' the predictor, after accounting for the measureable effects of all other predictors. Even if the classical regression model does not hold, e.g. when the response is ordinal, the partial correlation still serves as a heuristically useful benchmark for gauging the practical importance of an association.

A partial correlation can be calculated from the usual t-statistic used for testing the significance of a regression coefficient. Let T denote a t-statistic (the estimated coefficient divided by its standard error) and ν its degrees of freedom. The estimated partial correlation is

$$r = \pm\sqrt{\frac{T^2}{T^2+\nu}},$$

where the sign is chosen to be consistent with that of T. Under an assumption of multivariate normality, r is approximately nor-

Table 9.8. *Estimated partial correlation coefficients, 95% intervals and percent missing information from m = 10 multiple imputations under (a) data-dependent and (b) Jeffreys priors*

	(a) Data-dependent				(b) Jeffreys			
variable	est.	low	high	mis.	est.	low	high	mis.
LAN_2	−.08	−.21	.05	14	−.08	−.20	.05	14
LAN_3	.07	−.05	.20	12	.07	−.06	.20	11
LAN_4	−.10	−.34	.15	76	−.33	−.54	−.07	78
AGE_L	.11	−.03	.24	24	.04	−.15	.23	60
PRI_L	.24	.11	.37	20	.26	.11	.40	36
SEX_2	.04	−.10	.17	21	.03	−.12	.18	33
FLAS	.28	.14	.40	28	.29	.14	.42	40
MLAT	.18	−.01	.36	60	.22	.06	.36	40
SATV	−.06	−.21	.08	33	−.05	−.22	.13	51
SATM	.02	−.11	.15	15	.06	−.07	.19	15
ENG	−.03	−.18	.12	37	−.08	−.23	.07	36
HGPA	.45	.34	.55	20	.46	.35	.56	18
CGPA	.16	.02	.30	32	.14	.00	.27	21

mally distributed about the population coefficient ρ. An even better approximation is provided by Fisher's (1921) transformation $\tan^{-1}(r)$, which in large samples is essentially normally distributed about $\tan^{-1}(\rho)$ with variance $1/(\nu - 1)$ (Anderson, 1984).

For each imputed dataset, we regressed GRD on the same set of predictors used in the proportional-odds model. Using Rubin's rules, we calculated estimates and 95% intervals for $\tan^{-1}(\rho)$, and then transformed the results back to the correlation scale. The resulting point and interval estimates are shown in Table 9.8. These figures should be interpreted somewhat loosely, because the assumptions underlying the classical regression model and the normal approximation to $\tanh^{-1}(r)$ clearly do not hold. Yet it is apparent that FLAS, the predictor of primary interest, has substantial validity for predicting achievement in the study of foreign languages. Except for HGPA, FLAS has the highest partial correlation with GRD, higher even than the well established instrument MLAT.

9.5.3 National Health and Nutrition Examination Survey

The largest and most notable application of these methods to date has been to the Third National Health and Nutrition Examination

Survey (NHANES III). This survey, conducted by the National Center for Health Statistics, provides basic information on health and nutritional status for the civilian noninstitutionalized U.S. population. NHANES III is a complex, multistage area sample with oversampling of young children, the elderly, Mexican Americans and African Americans. Details of the design are given by Ezzati *et al.* (1992). Data were collected over six years (1988–94) with a total sample size of 39695. The data collection occurred in two stages: (a) personal interviews with subjects at home, and (b) detailed physical examinations of subjects in Mobile Examination Centers (MECs). Because of the inconvenience associated with going to a MEC and completing the exam, nonresponse rates at the examination phase were understandably high; many key survey variables had missingness rates of 30% or more.

In 1992, NCHS initiated a research project to investigate alternative missing-data procedures for NHANES III, including multiple imputation. This project will culminate in the public release of a multiply-imputed research dataset, currently scheduled for 1997. The dataset will contain five imputations of more than 60 variables. Here we briefly summarize the imputation model and the results of an extensive simulation study to assess the performance of the method. Complete details are given by Schafer *et al.* (1996) and their references.

The imputation model

The imputation model was designed to produce imputations appropriate for a wide variety of analyses. Data from NHANES are used to estimate important health-related quantities at the national level, e.g. rates of obesity by age and sex. These estimates, produced and reported by NCHS, are based on classical methods of survey inference (Cochran, 1977) and are designed to be approximately unbiased over repetitions of the sampling procedure. Standard errors are calculated using special variance-estimation techniques appropriate for data from complex samples (Wolter, 1985). To be compatible with these procedures, an imputation model must be sensitive to major features of the sample design. Outside NCHS, the data are also subjected to secondary analysis by researchers in many health-related fields. For example, researchers might fit linear or logistic regression models to NHANES data to investigate relationships among health outcomes and potential risk factors. For this reason, the imputation model needed to preserve [illegible]

marginal and conditional associations among variables.

We created multiple imputations under a general location model that included over 30 variables. Because individuals' probabilities of selection varied by age group, gender and race/ethnicity, the distributions of other survey variables had to be allowed to vary across the levels of these three; otherwise, biases could be introduced into many important estimators, both nationally and within demographic subclasses. The imputation model was also designed to reflect potential variation in characteristics across primary sampling units (PSUs), the clusters that enter into the NCHS procedures for variance estimation; without these effects, the quality of the standard errors calculated from the resulting imputed datasets could be impaired.

The categorical part of the general location model used a four-way classification by age, gender, race/ethnicity and PSU. The remaining variables were modeled by a multivariate linear regression with full three-way interactions for age, gender and race/ethnicity, plus main effects for PSUs. Most of the response variables in this regression were continuous, but a few were binary or ordinal. Multiple imputations were generated using the DABIPF algorithm of Section 9.4.4, and the imputed values for the binary and ordinal variables were rounded off to the nearest category. Preliminary analyses of the imputed data suggested that for most purposes, $m = 5$ imputations would be sufficient to obtain accurate and efficient inferences.

A simulation study

Recognizing that this imputation procedure was based upon a probability model that was, at best, only approximately true, we carried out an extensive simulation experiment. The goal of this simulation was to evaluate the performance of the imputation procedure from a purely frequentist perspective, without reference to any particular probability model. For example, we wanted to learn whether 95% interval estimates in typical applications would really cover the quantity of interest 95% of the time over repetitions of the sampling and imputation procedure. To this end, we constructed an artificial population of 31 847 persons by pooling data from four NCHS examination surveys conducted since 1971. This artificial population was weighted to resemble the projected U.S. population in the year 2000 in terms of race/ethnicity and geography. From [illegible] stratified random samples of 6000 persons

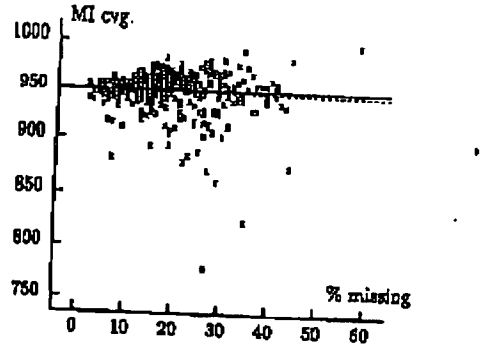


Figure 9.6. *Simulated coverage of 95% multiple-imputation (MI) intervals by average percent missing information for 448 means.*

using a sampling plan resembling that of NHANES III. Missing values were imposed on each sample using a random, ignorable mechanism to mimic the rates and patterns of nonresponse observed in NHANES III. The missing data were then imputed five times under a general location model, and multiple-imputation point and interval estimates were calculated for a variety of estimands (means and proportions, subdomain means, quantiles, and conditional log-odds ratios) using methods appropriate for stratified random samples. The entire sampling, imputation and estimation procedure was repeated 1000 times.

Here we briefly summarize our results for means. We examined means for ten exam variables for the entire population and within demographic categories defined by age, race/ethnicity and gender. Among these 448 means, the average simulated coverage of the 95% intervals over 1000 repetitions was 949.3, not significantly different from 950. Individually, however, 81 of the 448 means (18%) had coverage significantly different from 950 at the 0.05 level. The coverages of the multiple-imputation (MI) intervals are shown in Figure 9.6, plotted against the average estimated percent missing information for the respective estimands. In this plot, the least-squares fit (dashed line) is nearly indistinguishable from a horizontal line through 950 (solid); there is no overall tendency for the actual coverage to increase or decrease with the fraction of missing information. There is, however, some tendency for the coverage to vary more as the rate of missing info[illegible]

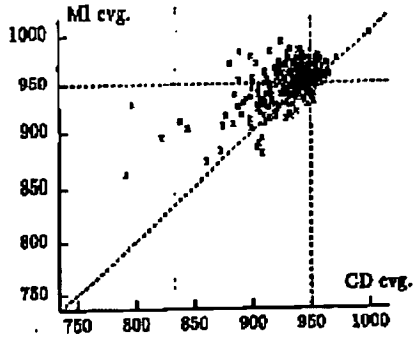


Figure 9.7. *Simulated coverage of 95% multiple-imputation (MI) intervals versus complete data (CD) intervals, with points (507, 824), (608, 799) and (479, 876) not shown.*

Further analysis revealed that, among the intervals whose coverage departed substantially from 95%, the departures could be largely traced to failure in the normal approximation for the inference without missing data. In Figure 9.7, the simulated coverage of each MI interval is plotted against the coverage of the corresponding normal-based interval (the point estimate plus or minus 1.96 standard errors) that one would have used if no data were missing. The two coverages are strongly correlated. Somewhat surprisingly, for the estimands for which the complete-data (CD) interval exhibited gross undercoverage (and especially the three pathological cases that fell outside the plotting region) the MI intervals performed substantially better than their CD counterparts. On the other hand, there were no estimands for which CD did well but MI did poorly. Results for other types of estimands revealed similar trends: the MI intervals tended to perform very well, except where difficulties were observed in the corresponding CD intervals. Further discussion of this simulation study, including its limitations, are given by Schafer *et al.* (1996).

Further remarks

In this application, it was feasible to add PSU to the general location model because there were relatively few PSUs and a large number of subjects within each PSU; we were able to include dummy

lems of inestimability. In other surveys, the number of clusters may be too large to adopt such an approach. In those settings, it may be possible to produce multiple imputations under hierarchical or random-effects models that impose probability distributions on the cluster-specific parameters. Estimation and imputation algorithms for random-effects models can be developed by extending the techniques of this chapter, but they are beyond the scope of this book. For an example of imputation under a random-effects model for multivariate categorical data, see Schafer (1995).